numero 154
giugno 1981
anno XIV
volume XXVI

LE SCIENZE
edizione italiana di SCIENTIFIC AMERICAN

Emissioni accidentali di radioattività

L'incidente più grave immaginabile per un reattore nucleare è di gran lunga meno distruttivo dell'esplosione di un'arma nucleare, anche nell'ipotesi che questa causi danni solo mediante radiazioni

di Steven A. Fetter e Kosta Tsipis

L'esposizione a una quantità pericolosa di radioattività è un evento che per una popolazione numerosa potrebbe verificarsi in circostanze diverse. Ad esempio in seguito a una guerra nucleare, anche la più circoscritta, oppure in caso di incidente a un reattore nucleare, che provochi la rottura del recipiente di contenimento, con conseguente fuoriuscita di materiale dal nocciolo del reattore ed emissioni di sostanze radioattive nell'atmosfera; di minore entità sarebbe invece il pericolo rappresentato da fortuite fuoriuscite di acqua e gas radioattivi da un reattore. È inoltre possibile che si verifichino incidenti nel corso delle operazioni di fabbricazione, trasporto, ritrattamento e immagazzinamento di materiale radioattivo destinato a reattori o ad armi nucleari.

Poiché l'entità delle possibili emissioni di radioattività varia nettamente secondo il tipo di evento considerato, è necessario esaminare separatamente ciascun evento. Prendiamo qui in considerazione tre cause distinte di possibili emissioni di radioattività: l'esplosione al suolo di un'arma termonucleare; la fusione del nocciolo di un reattore nucleare, con rottura del recipiente di contenimento e conseguente fuga di radioattività; l'esplosione di una testata termonucleare contro un reattore nucleare.

È bene precisare che in questi confronti non si tiene conto né dell'onda di pressione né del calore, che rappresentano gli effetti immediati ed esplosivi di un'arma termonucleare, ma ci si limita a esaminare e a confrontare i soli effetti tardivi dell'emissione di radioattività. Tuttavia, l'esplosione di un'arma nucleare risulta molto più temibile di qualsiasi incidente a un reattore; inoltre, l'esplosione di una testata nucleare contro un reattore è nettamente più distruttiva in confronto alla sola esplosione al suolo. L'attacco nucleare trasforma il reattore in un'arma radiologica devastatrice.

Un'arma termonucleare è costituita normalmente da tre parti. La prima è essenzialmente un dispositivo di innesco, il cui componente principale è plutonio nella quantità di alcuni chilogrammi. La fissione dei nuclei di plutonio produce il calore necessario per l'innesco dell'esplosione termonucleare.

La seconda parte è rappresentata dall'esplosivo termonucleare, una miscela di deuterio e trizio, gli isotopi pesanti dell'idrogeno. Dalla fusione termonucleare di un nucleo di deuterio (che ha un protone e un neutrone) con un nucleo di trizio (un protone e due neutroni) si produce un nucleo di elio (due protoni e due neutroni). Il neutrone in più viene emesso ad alta velocità e viene liberata sotto forma di calore una certa quantità di energia. I prodotti di tale reazione sono caratterizzati da una radioattività di durata relativamente breve.

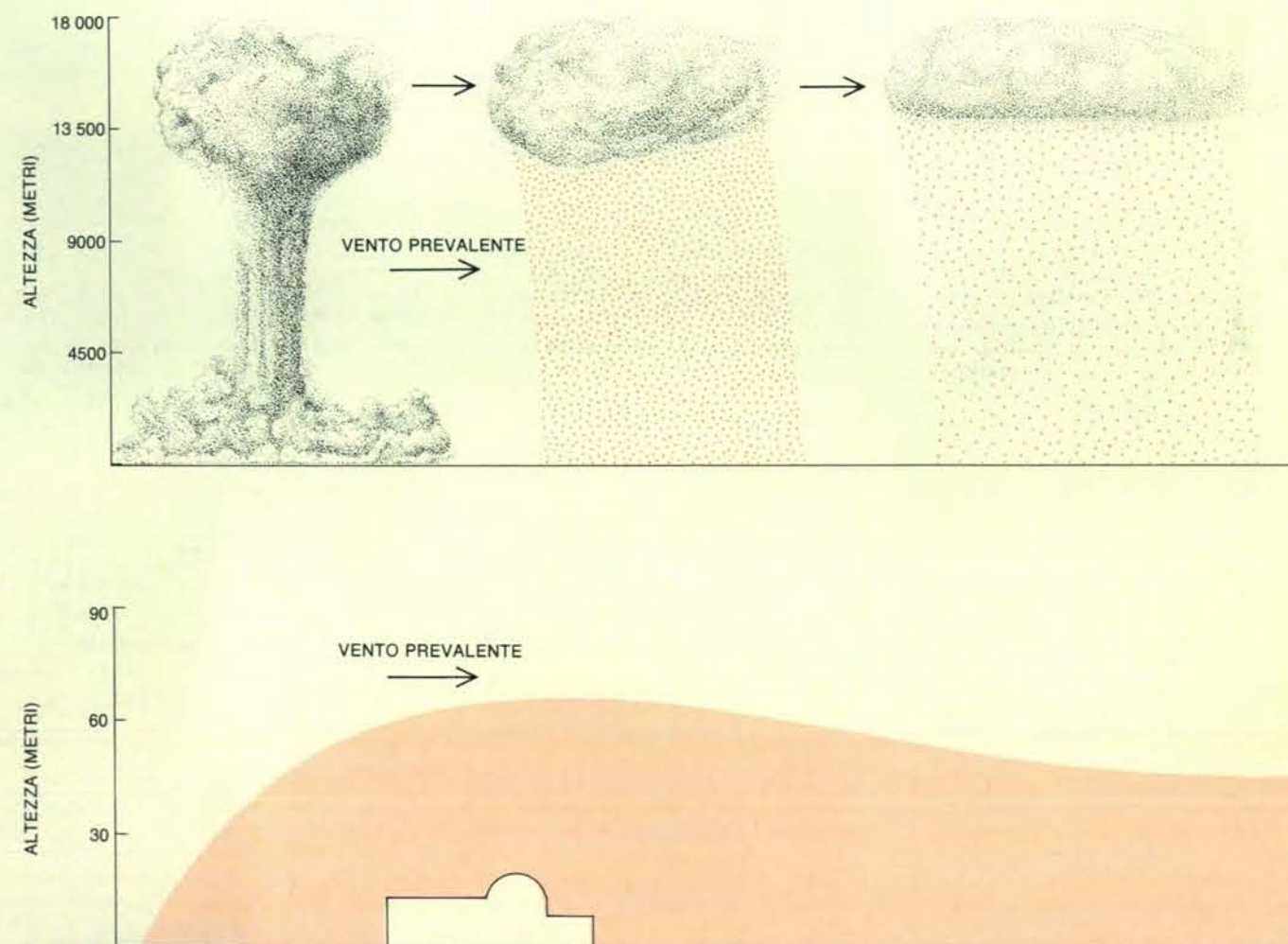
La terza parte dell'arma è un mantello di uranio, che circonda lo strato di deuterio e trizio. I nuclei degli atomi di uranio subiscono fissione in seguito al bombardamento dei neutroni emessi durante la fusione termonucleare; i frammenti dei nuclei fissionati sono intensamente radioattivi. In un'arma termonucleare a tre parti, del tipo descritto, circa la metà dell'energia liberata proviene dalla fusione termonucleare e l'altra metà dalla fissione dell'uranio.

All'atto dell'esplosione, il calore generato provoca la quasi istantanea volatilizzazione dell'arma termonucleare, e quindi l'arresto delle reazioni nucleari. I nuclei generati dalla fissione dell'uranio si trovano quasi tutti in uno stato energetico anomalo e il loro ritorno allo stato di energia inferiore si accompagna all'emissione di radiazione nella parte ad alta energia dello spettro elettromagnetico, quella parte rappresentata dalle radiazioni X e gamma. Questa radiazione riscalda l'aria circostante, costituendo un'onda d'urto, che a sua volta riscalda ulteriori strati di aria. Il risultato di tutto ciò è una luminosa sfera di fuoco. Per un'arma con un potenziale esplosivo di un megaton (l'equivalente in energia di un milione di tonnellate di esplosivo chimico) la sfera di fuoco sale, con una velocità dell'ordine di 120 metri al secondo, a un'altezza di circa 18 000 metri. Il potenziale di un megaton è quello tipico delle testate dei missili balistici intercontinentali dell'Unione Sovietica.

La corrente ascensionale generata dalla sfera di fuoco solleva una gran quantità di terreno e detriti. Un'esplosione da un megaton al suolo è in grado di scavare un cratere di 120 metri di profondità e di 360 metri di diametro. Man mano che la sfera di fuoco si raffredda, i nuclei radioattivi generati dall'esplosione contaminano le particelle di polvere in sospensione nell'aria, le quali nel corso di un certo periodo di tempo ritornano a terra sotto forma di ricadute radioattive.

Le ricadute sono in parte radioattive perché alcuni dei nuclei di nuova generazione sono instabili, normalmente per un eccesso di neutroni. L'instabilità viene a ridursi con la trasformazione di un neutrone in un protone, attraverso il processo noto come decadimento beta; nel corso di questo processo il nucleo emette un elettrone, che in questo contesto prende il nome di raggio beta. Dopo queste trasformazioni il nucleo può trovarsi ancora in uno stato eccitato, dal quale passa a un livello energetico inferiore con l'emissione di radiazione elettromagnetica, essenzialmente radiazione gamma. Le particelle di ricaduta continuano a emettere radiazione beta e gamma per diversi decenni dall'esplosione, e in modo casuale. Dato un certo numero di nuclei radioattivi, si può prevedere solo il numero medio di emissioni in un dato intervallo temporale. Col passare del tempo si riduce il numero dei nuclei in stati eccitati o instabili e quindi l'intensità della radioattività viene a ridursi.

Per definire la quantità di radioattività, o la quantità di energia che la radioattività può depositare in un tessuto vivente, si usano varie unità di misura. L'unità standard di misura della radioattività vera e propria è il curie, definito come $3,7 \times 10^{10}$ emissioni al secondo. Questa definizione è svincolata sia dal tipo di radiazione sia dall'energia a questa associata. Per quanto riguarda l'energia depositata dalla radioattività, la corrispondente unità di misura è il rad, definito come l'assorbimento di 100 erg da parte di un grammo di materiale equivalente al tessuto vivente. Un'altra unità è il roentgen, che riguarda esclusivamente la radiazione gamma e X. L'esposizione a un roentgen di radiazione gamma equivale all'assorbimento di circa 94 erg da parte di un grammo di tessuto. Il rad e il roentgen sono quindi unità di misura praticamente equivalenti.

Occorre però ancora un'altra unità di misura, in quanto nessuna di quelle citate consente di descrivere l'entità del danno biologico causato dalla radiazione. Tale unità è il rem, abbreviazione di *roentgen equivalent man*. La misura di dose di radiazione in rem tiene conto del fatto che radiazioni di natura diversa possono dare luogo a effetti diversi su un organismo vivente, anche se la quantità di energia depositata è la stessa e anche se il meccanismo generale del danneggiamento ha la stessa causa, ossia la ionizzazione degli atomi presenti nelle molecole intracellulari. Le differenze nell'entità del danneggiamento riflettono le differenze nelle caratteristiche delle radiazioni, come ad esempio la diversa lunghezza del percorso entro il tessuto. Una dose in rem equivale a una dose in rad moltiplicata per un fattore deno-

La nube dei prodotti dell'esplosione di un'arma nucleare è diversa dalla nube causata dall'incidente a un reattore. L'esplosione di una testata termonucleare da un megaton al suolo *(disegno in alto)* provoca una corrente ascensionale che solleva i prodotti dell'esplosione a un'altezza di circa 18 000 metri; tali prodotti, trasportati dal vento, ricadono a terra come pulviscolo radioattivo. Invece, all'esplosione di tipo non nucleare che provochi la fessurazione del recipiente di contenimento del reattore *(disegno in basso)* è associata un'energia relativamente bassa, per cui la radioattività emessa non viene portata ad alta quota. La dispersione di radioattività da parte del vento prevalente è inoltre ridotta.

minato efficacia biologica relativa (EBR) del particolare tipo di radiazione. Per le radiazioni beta e gamma EBR vale circa uno; nel seguito considereremo, quindi, equivalenti le dosi espresse in rad e in rem. Come concreto termine di riferimento per la dose di radiazione valgano i seguenti dati: una radiografia al torace comporta una dose di circa 0,01 rem, assorbita in una frazione di secondo, mentre la radiazione naturale di fondo a livello del mare corrisponde a circa 0,075 rem all'anno.

Gli effetti biologici della radiazione variano considerevolmente da individuo a individuo; essi dipendono, ad esempio, dall'età e dallo stato di salute del soggetto. Non è pertanto possibile indicare con precisione i livelli di radiazione in corrispondenza dei quali è prevedibile il manifestarsi di sintomi caratteristici della sindrome da radiazioni, quali caduta dei capelli, vomito, diarrea, emorragie o lesioni interne alla bocca e alla gola. È stato però dimostrato che la sopravvivenza è praticamente impossibile se il corpo umano subisce un'esposizione superiore a 500 o 600 rem per un intervallo di tempo non superiore a uno o due giorni. Se la dose è compresa tra 200 o 450 rem la sopravvivenza è possibile, ma non certa anche se si effettuano le dovute terapie mediche. Si può ragionevolmente affermare che una dose di 400 rem in un giorno comporti un tasso di mortalità del 50 per cento o più. Per una popolazione, un'esposizione di 100 rem comporterebbe casi di infermità e alcuni casi letali, ma la maggior parte riuscirebbe prevedibilmente a guarire anche senza terapie mediche specialistiche.

Nel calcolo dell'estensione di territorio reso inagibile da una emissione di radioattività, consideriamo che la massima dose ammissibile per la popolazione in genere sia di 2 rem annui. Questa dose supera di oltre dieci volte quella massima prescritta dalla US Environmental Protection Agency, e di oltre 20 volte quella corrispondente alla radiazione naturale di fondo. D'altra parte, questo livello è inferiore ai 5 rem annui attualmente considerati come limite superiore per i lavoratori professionalmente esposti alle radiazioni per anni. In tempo di pace, questo limite potrebbe benissimo venire ridotto a 2 rem annui: il provvedimento potrebbe, ad esempio, essere deciso in seguito a un eventuale incidente a un reattore. Ma in caso di guerra nucleare è ben difficile pensare che la popolazione possa completamente essere allontanata dalle zone in cui la radiazione superi il livello di 2 rem annui. Anzi, spinta dalla fame o da altre necessità, la popolazione potrebbe essere disposta (o costretta) a occupare zone in cui il livello superi anche i 50 rem annui; un livello, questo, in corrispondenza del quale si registrano sindromi da radiazione in oltre il 50 per cento dei soggetti esposti. Una dose di 50 rem annui sporadicamente può essere letale e in qualche caso può provocare l'insorgere del cancro a distanza di anni dall'esposizione.

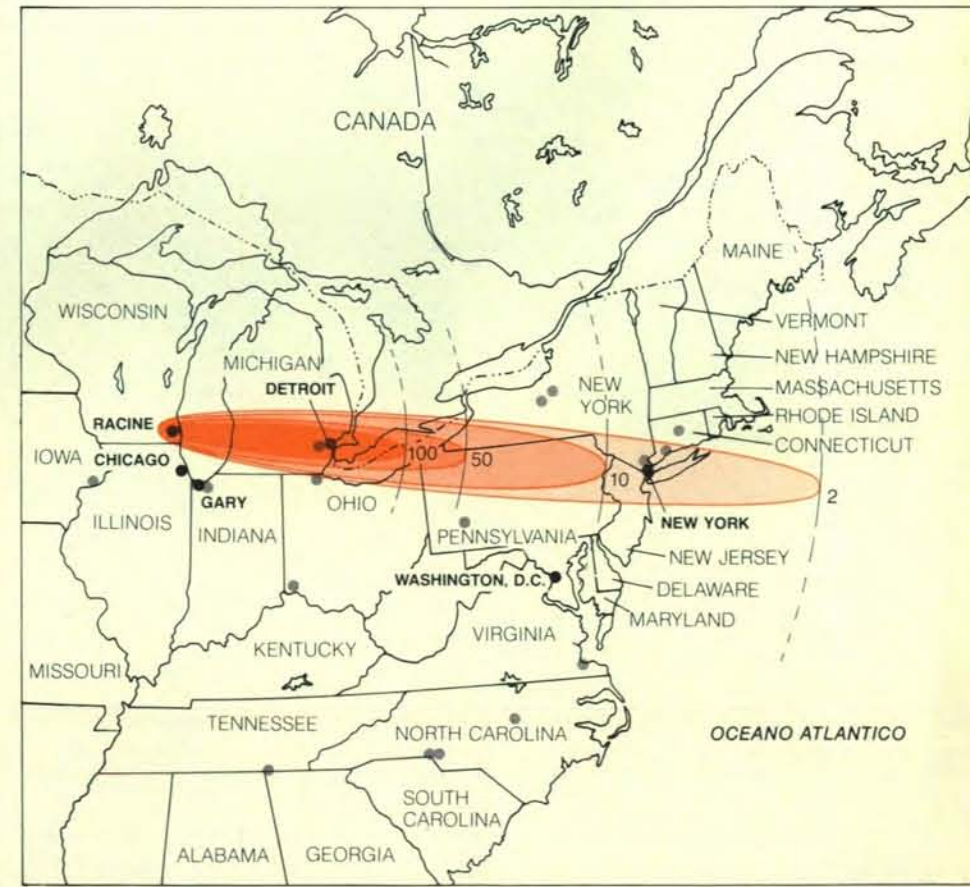
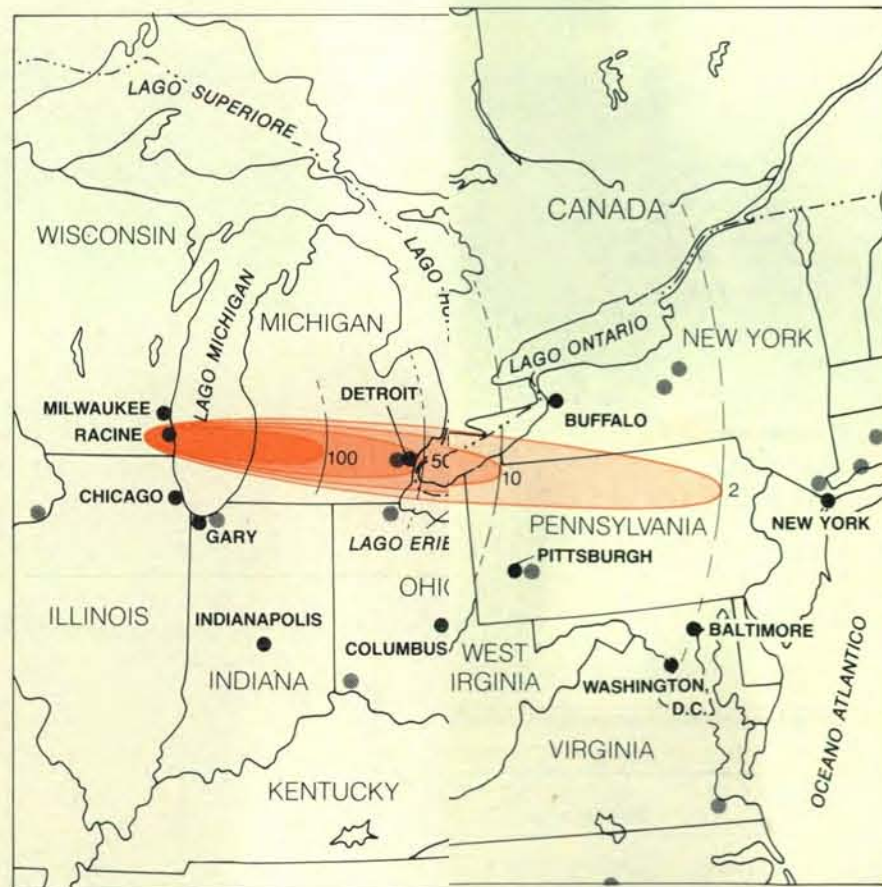
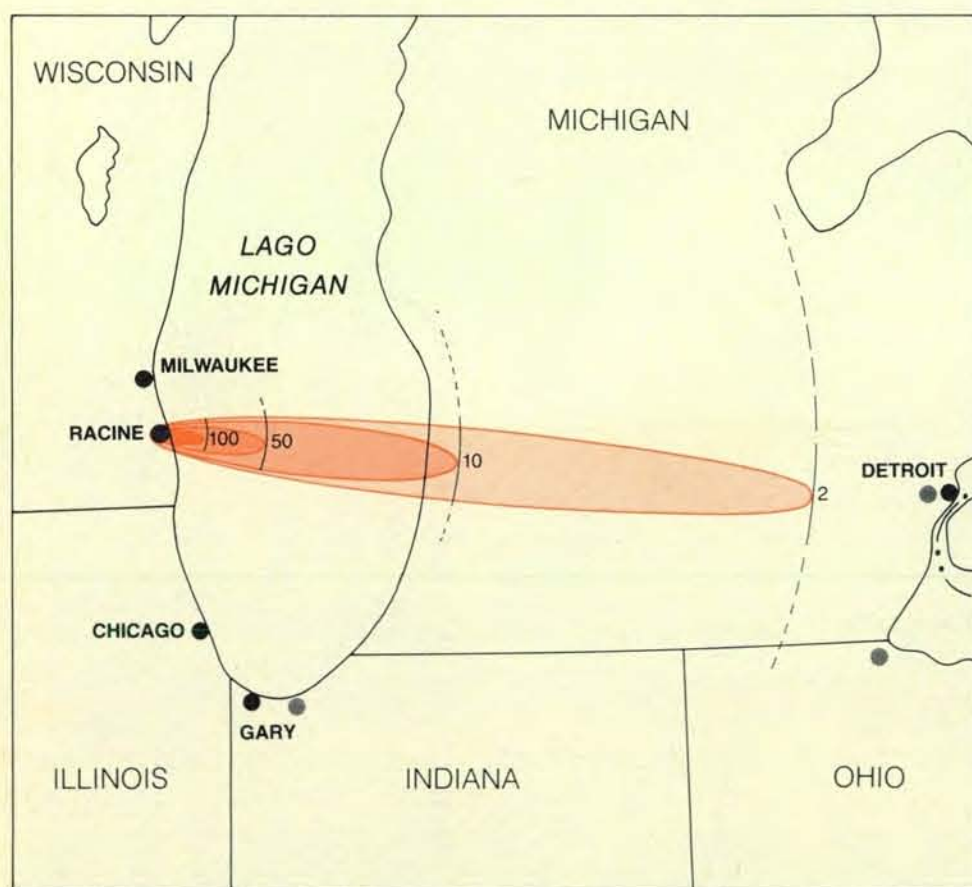
Ritorniamo ora alle conseguenze dell'emissione di radioattività da parte di un'arma termonucleare da un megaton, fatta esplodere a livello del suolo. La maggior parte del pulviscolo radioattivo prodotto ricade sottovento subito dopo l'esplosione; circa il 70 per cento di questo pulviscolo è costituito da particelle relativamente grandi, che ricadono a terra nel giro di una giornata. L'intensità della radioattività decresce con la distanza dal luogo dell'esplosione, innanzitutto perché la nube dei prodotti d'esplosione viene a perdere particelle di pulviscolo spostandosi con il vento, e poi perché la radioattività globale si riduce in seguito al decadimento dei nuclei radioattivi.

Se il vento è costante, la distribuzione della dose accumulata di radioattività (in rem) è rappresentata da un insieme di profili sigariformi di livello che si intersecano. Ogni profilo indica una dose particolare e tutti i punti al suo interno presentano valori più elevati. Avanziamo l'ipotesi che la velocità del vento sia di 24 chilometri all'ora. In questo caso la zona letale - ossia l'area delimitata dal profilo che corrisponde a un'esposizione di 400 rem in 24 ore - misura circa 1024 chilometri quadrati. Il numero dei morti nella zona letale sarebbe strettamente legato alla densità della popolazione. Negli Stati Uniti, ad esempio, la densità varia da 39 000 persone per chilometro quadrato nelle ore lavorative a poco meno di due persone per chilometro quadrato. Quindi, la radiazione prodotta dall'esplosione di una singola testata nucleare potrebbe causare la morte di un numero di persone compreso tra alcune centinaia di unità e diversi milioni. La cifra definitiva dipenderebbe non solo dal luogo dell'esplosione, ma anche dall'ora, dalle condizioni atmosferiche, dalla maggiore o minore efficacia delle azioni preventive, oltre che dalla disponibilità o meno di mezzi per la protezione contro le radiazioni.

I superstiti non potrebbero rientrare nella zona letale per un lungo periodo di tempo, a causa della contaminazione radioattiva del terreno, e sarebbero costretti ad attendere che gli effetti del decadimento radioattivo, unitamente a quelli della penetrazione nel terreno del materiale contaminante a opera delle precipitazioni atmosferiche, portino alla riduzione della radioattività entro livelli accettabili. Se si considera come massimo ammissibile un livello di 2 rem annui, circa 3070 chilometri quadrati di territorio resterebbero inagibili per un anno, mentre aree più vaste sarebbero inagibili per periodi inferiori. L'impatto sociale sarebbe gravissimo. Ad esempio, per un mese risulterebbe inagibile una zona di oltre 51 000 chilometri quadrati, e ciò comporterebbe lo sfollamento di varie centinaia di migliaia di persone.

Un attacco con diverse armi nucleari causerebbe un accumulo di radioattività talmente elevato che quasi certamente non sarebbe più ipotizzabile, per i superstiti, il ripristino dell'attività produttiva nemmeno nelle zone indenni dall'esplosione. Anche se fossero disposti a insediarsi in aree con corrispondente dose di esposizione nettamente superiore ai 2 rem annui, ai superstiti resterebbero ancora precluse vaste zone di territorio. Ciascuna arma da un megaton dà luogo a una zona di 3840 chilometri quadrati in cui la dose di radiazione si mantiene almeno a 50 rem annui per un mese.

Contrariamente a un'arma nucleare, un reattore nucleare non può esplodere. L'energia viene liberata dal reattore attraverso la fissione nucleare, ma anche in un reattore che per ipotesi fosse completamente senza controllo il tasso di emissione dell'energia è oltre 10^{12} volte inferiore al corrispondente tasso di emissione di un'arma nucleare. Inoltre, l'energia prodotta nel reattore viene assorbita inizialmente dalla massa del nocciolo, che è centinaia di volte più grande della massa di un'arma nucleare. Ne consegue che, anche in un reattore privo di controllo, la temperatura del nocciolo cresce solo lentamente.

Se la temperatura del nocciolo crescesse a dismisura, gli elementi di combustibile fonderebbero e il nocciolo collasserebbe prima che una reazione a catena desse luogo a una quantità di energia in grado di provocare un'esplosione. In seguito,

Confronto tra profili di dose risultanti da tre ipotetiche emissioni di radioattività. Nella mappa a sinistra è rappresentata la situazione a una settimana di distanza da un incidente in cui dal nocciolo di un reattore nucleare da un gigawatt venga emesso un terzo della radioattività totale presente. L'ipotetica località dell'incidente è Racine, nel Wisconsin. La quantità di radioattività emessa ammonta a 100 milioni di volte quella liberata dalla centrale nucleare di Three Mile Island, presso Harrisburg, Pennsylvania, nel marzo 1979. Nella mappa al centro, riferita sempre alla città di Racine, è raffigurata la situazione a una settimana di distanza dall'ipotetica esplosione di un'arma nucleare al suolo, senza danni al reattore nucleare esistente, e con una emissione iniziale di radioattività di gran lunga superiore rispetto all'esempio precedente. La mappa sulla destra rappresenta la situazione a una settimana di distanza dall'esplosione di un'arma nucleare da un megaton, che provochi anche la simultanea volatilizzazione del nocciolo di un reattore da un gigawatt. Sia la radioattività emessa dal reattore sia quella emessa dall'arma nucleare si disperdono sul territorio. Per tutti i casi considerati il vento prevalente soffia da ovest alla velocità di 24 chilometri all'ora. La nube dei prodotti dell'esplosione può però dirigersi in direzioni diverse *(pallini in grigio più grandi)*. Le dosi sono espresse in rem annui. Un rem corrisponde a una quantità di radiazione che depositi un'energia di 100 erg in un grammo di tessuto vivente. La radiazione naturale di fondo al livello del mare corrisponde a circa 0,075 rem annui. L'esposizione per un anno di una popolazione a 2 rem potrebbe aumentare a lungo termine i casi di cancro. L'esposizione a 50 rem in un anno può causare sindrome da radiazioni. I puntini in grigio indicano i reattori nucleari in esercizio o in costruzione, a non più di 40 chilometri da una città con oltre 100 000 abitanti.

un'eventuale rottura nel recipiente di contenimento potrebbe quindi portare a un'emissione di radioattività. Si può infatti ipotizzare un incidente in cui la perdita totale del refrigerante delle barre di combustibile porti le barre stesse al surriscaldamento e alla fusione. Il materiale fuso entra perciò in contatto con l'acqua e un'esplosione di vapore porta allo scoperchiamento del recipiente di contenimento, con conseguente fuoriuscita di materiale radioattivo. Si può immaginare un altro tipo di incidente, in cui il surriscaldamento del nocciolo porti alla generazione di idrogeno o di altri gas infiammabili, che miscelandosi con l'ossigeno atmosferico provochino un incendio e una conseguente esplosione. Anche in questo caso si avrebbe la rottura del recipiente di contenimento e una fuga di radioattività.

Per un confronto tra i rischi connessi all'emissione di radioattività da un reattore e da un'arma nucleare, esaminiamo le conseguenze del più grave incidente ipotizzabile, quello cioè in cui si verifichi la rottura del recipiente di contenimento. È bene mettere in evidenza che la probabilità calcolata di un simile incidente è di molti ordini di grandezza inferiore alla probabilità di un incidente più lieve, del tipo di quello verificatosi nel marzo 1979 alla centrale nucleare di Three Mile Island, presso Harrisburg, nello stato di Pennsylvania.

La quantità di materiale radioattivo che può fuoriuscire da un reattore, e la composizione di questo materiale, dipendono rispettivamente dalle particolari modalità dell'incidente e dal periodo di tempo trascorso dall'ultima ricarica di combustibile. La dispersione della radioattività può dipendere dalla forma della nube di prodotti fuoriusciti nell'incidente e dalle condizioni atmosferiche locali. Si possono trarre due conclusioni di carattere generale. In primo luogo, il tasso di emissione di radioattività da un'arma termonucleare è inizialmente di gran lunga più elevato di quello corrispondente a un incidente di reattore, anche se la radioattività prodotta dall'arma nucleare è dovuta per lo più a isotopi con breve vita radioattiva. In secondo luogo, un incidente a un reattore dà luogo a una quantità di calore relativamente modesta. Ne consegue che la nube dei prodotti fuoriusciti non sale a grande altitudine e lascia ricadere la propria radioattività in breve tempo. Le dimensioni della zona contaminata sono così limitate. In conclusione, l'area di territorio contaminata in seguito a un incidente di reattore è assai più ridotta, ma la contaminazione vi persiste più a lungo.

L'attacco a un singolo reattore con una singola arma nucleare potrebbe causare la devastazione di una parte cospicua dell'Europa. Nell'illustrazione si considera un ipotetico attacco al reattore nucleare da un gigawatt di Neckarwestheim, nella Germania Occidentale. L'arma nucleare ha carica esplosiva di un megaton. Il vento prevalente soffia da sud-est con una velocità di 24 chilometri all'ora. A un mese di distanza dall'attacco la zona in cui la dose aumenta a 10 rem annui *(in colore chiaro)* potrebbe estendersi senza difficoltà alla Gran Bretagna. A un anno di distanza la zona corrispondente a una dose di 10 rem *(in colore scuro)* interesserebbe ancora la maggior parte delle regioni industriali della Germania Occidentale. I punti in grigio indicano i siti dei reattori nucleari di potenza della Germania Occidentale e della Francia.

Consideriamo un reattore nucleare della potenza di un gigawatt (1000 megawatt), nel quale ogni anno viene effettuato il ricambio di un terzo del combustibile, e immaginiamo che in seguito a un'esplosione di vapore il recipiente di contenimento si fessuri e nell'atmosfera vengano emessi un terzo di tutti i nuclei radioattivi presenti nel reattore. Dopo un'ora la radioattività del materiale fuoriuscito ammonterebbe a circa 1,5 miliardi di curie. L'esplosione di un'arma nucleare da un megaton darebbe luogo a una quantità di radioattività 1000 volte superiore. Nell'incidente di Three Mile Island ne venne liberata una quantità 100 milioni di volte inferiore (e precisamente 17 curie di iodio radioattivo).

Ipotizziamo che la velocità del vento sia di 24 chilometri all'ora. Le conseguenze tipiche dell'incidente considerato sono legate all'estensione relativamente ridotta dell'area contaminata. Inoltre, il livello di esposizione delle radiazioni si mantiene prossimo ai 2 rem annui in tutta la zona contaminata, con l'eccezione di un'area ristretta. Più precisamente, la dose si mantiene sul valore di 2 rem annui per un mese in un'area di circa 4600 chilometri quadrati. (L'area corrispondente, per l'esplosione di un'arma nucleare da un megaton, è di 51 200 chilometri quadrati.) La zona letale, ove la dose raggiunge il livello di 400 rem al giorno, ha un'estensione inferiore a 2,5 chilometri quadrati (e di 1025 chilometri quadrati per un'arma nucleare da un megaton).

In seguito a un incidente a un reattore nucleare, per i ridotti valori di dose e di estensione dell'area contaminata, la popolazione potrebbe venire evacuata prima che l'inalazione di pulviscolo radioattivo sia significativa, evitando così il principale pericolo derivante da un simile incidente. È anche verosimile che il territorio possa essere decontaminato. Nel caso dell'arma nucleare, la decontaminazione del territorio sarebbe invece preclusa dall'assai più elevata quantità di radioattività ricaduta.

In questo confronto tra possibili danni provocati da un incidente a un reattore e dall'esplosione di un'arma nucleare è opportuno considerare gli eventi distruttivi di tipo immediato. La fusione del nocciolo di un reattore non provoca conseguenze dannose rilevanti per lo scoppio dell'involucro o per il calore, mentre l'esplosione dell'arma nucleare comporta l'immediata devastazione della zona entro un raggio compreso tra 8 e 16 chilometri. Con ogni probabilità nel caso dell'arma nucleare verrebbero distrutti e messi fuori uso i servizi sanitari e di emergenza. Si può quindi concludere che una popolazione esposta a una dose pericolosa di radiazioni derivante dall'esplosione di un'arma nucleare avrebbe probabilità di sopravvivenza molto inferiori rispetto a una popolazione esposta alla stessa dose, ma che derivasse da un incidente a un reattore, e questo proprio a causa della distruzione dei servizi essenziali per le vittime dell'esposizione alla radiazione.

È difficile valutare la probabilità dei due eventi qui considerati. Tuttavia, sia gli strateghi militari sia gli esperti nucleari sembrano concordi nell'indicare che la probabilità che un'arma nucleare venga fatta esplodere in qualche parte del mondo entro il prossimo decennio è nettamente

L'estensione di territorio inagibile per i superstiti di una emissione di radioattività dipende dalla dose di radiazione che gli stessi superstiti sono in grado (o costretti) ad assorbire. È verosimile che in tempo di pace anche un'esposizione a dosi di pochi rem annui non sia ammissibile, mentre invece i superstiti di un attacco nucleare tenterebbero di resistere a condizioni molto più critiche. Gli istogrammi indicano l'estensione del territorio che deve restare disabitato per un anno se il tasso di dose ammissibile è rispettivamente di 10 rem annui *(a sinistra)*, di 50 rem annui *(al centro)*, o di 100 rem annui *(a destra)*. Anche qui vengono prese in considerazioni tre possibili cause di contaminazione radioattiva: un grave incidente a un reattore *(in colore chiaro)*, un'esplosione al suolo di un'arma termonucleare *(in colore intermedio)*, e l'esplosione di un'arma termonucleare da un megaton contro un reattore nucleare *(in colore scuro)*. Se un livello di 10 rem annui non fosse ammissibile, l'estensione di territorio che dovrebbe restare disabitata per un anno dal giorno dell'attacco al reattore avrebbe una superficie di 22 000 chilometri quadrati.

La radioattività emessa dall'esplosione di un'arma nucleare decade in modo diverso dalla radioattività dovuta a un incidente a un reattore, perché i vari isotopi sono presenti, nei due casi, in proporzioni diverse. Dopo un'ora, la radioattività liberata da un'arma nucleare da un megaton è 1000 volte più elevata di quella liberata nel più grave incidente a un reattore nucleare, ipotizzabile in tempo di pace. La radioattività emessa dal reattore decade, però, più lentamente. L'unità di radioattività, il curie, corrisponde a 3,7 × 10^{10} emissioni al secondo di vari tipi di radiazioni.

più elevata della probabilità della fusione catastrofica del nocciolo di un reattore. A supporto di questa indicazione stanno il numero sempre crescente di armi nucleari negli arsenali di diverse nazioni, l'irrigidimento delle relazioni tra Stati Uniti e Unione Sovietica, e la tendenza dei pianificatori militari a influenzare la politica strategica del proprio paese, imperniata sul concetto di guerra nucleare come deterrente, modificandola in una politica di preparazione a una tale guerra.

Consideriamo ora la radioattività che verrebbe emessa nel caso di esplosione di un'arma nucleare contro un reattore nucleare da un gigawatt, nell'ipotesi che il materiale radioattivo nel nocciolo del reattore venisse completamente volatilizzato nell'esplosione. In questo caso la radioattività del nocciolo si sommerebbe a quella prodotta dall'arma nucleare stessa, per innalzarsi con la sfera di fuoco e quindi ricadere a terra, secondo il processo caratteristico delle ricadute associate all'esplosione di armi nucleari.

Nella prima settimana l'andamento della contaminazione sarebbe sostanzialmente identico a quello corrispondente all'esplosione della sola arma nucleare, in quanto la radioattività emessa dal reattore è inizialmente più ridotta rispetto a quella emessa dall'arma. Poiché però la radioattività derivante dal reattore ha una vita relativamente lunga, il periodo di tempo in cui una zona resterebbe contaminata è sensibilmente più lungo. In sostanza, i prodotti dell'esplosione dell'arma nucleare sono responsabili dell'alto livello di radioattività iniziale, e quelli fuoriusciti dal reattore, della radioattività di lunga durata. La zona letale, nel caso considerato, supererebbe i 1280 chilometri quadrati, un'area superiore di un terzo rispetto alla zona letale creata dall'esplosione della sola arma nucleare. La zona in cui la dose cumulativa si manterrebbe per un mese al livello di 2 rem annui misurerebbe circa 164 000 chilometri quadrati, ossia sarebbe tre volte più estesa. La dose si manterrebbe poi al livello di 2 rem annui per un anno in una zona di 64 000 chilometri quadrati, cioè venti volte più estesa. Esisterebbe inoltre una zona di 460 chilometri quadrati, nella quale si sarebbe esposti a una dose di almeno 2 rem annui per più di un secolo. Tale zona rappresenterebbe il monumento permanente a una simile catastrofe.

È chiaro che la volatilizzazione del nocciolo di reattori nucleari rappresenta un efficace mezzo di devastazione di ampie regioni di un paese. Per la verità, anche con il solo impiego di un'arma termonucleare sarebbe possibile distruggere una frazione rilevante della potenza industriale di un paese nemico, purché in presenza di condizioni meteorologiche favorevoli. Ad esempio, un attacco contro un reattore nella valle del Reno potrebbe rendere inabitabile circa un terzo dei 245 000 chilometri quadrati della Germania Occidentale per un periodo di un mese o più, anche se i superstiti fossero disposti ad assorbire dosi cumulative di radioattività molto superiori ai 2 rem annui. L'unica condizione per il successo di un simile attacco sarebbe un vento prevalente che soffi da sud-est.

Quando si pensa a distruzioni di questa entità, è bene ricordare che nell'Europa centrale, dove la densità di popolazione è elevata e il terreno intensamente sfruttato, i reattori nucleari di potenza possono trovarsi a breve distanza da installazioni militari. Diventa quindi non trascurabile la probabilità che un'arma nucleare destinata a un obiettivo militare vada invece a distruggere un reattore nelle vicinanze. È anche opportuno ricordare che le vasche di deposito dei rifiuti radioattivi si trovano nello stesso sito del reattore da cui tali rifiuti provengono. E i rifiuti contenuti in una vasca tipica potrebbero in breve rappresentare un accumulo di radioattività due volte maggiore di quello del nocciolo. Inoltre, spesso i reattori vengono costruiti a coppie poco distanziate. Tutto sommato, i valori di dose associati all'esplosione di un'arma nucleare contro un complesso di reattori nucleari potrebbero benissimo essere da due a sei volte più elevati di quelli che noi abbiamo calcolato.

Non siamo in grado di affermare se gli strateghi militari abbiano considerato attentamente nei loro scenari di guerre nucleari il caso di deliberata o fortuita volatilizzazione del nocciolo di un reattore durante un attacco con testate nucleari. Il modo migliore di ridurre al minimo la probabilità di un simile evento sarebbe quello di evitare qualsiasi tipo di guerra nucleare. Sarebbe anche opportuna la definizione di un accordo multinazionale in base al quale le installazioni nucleari civili venissero escluse dal novero dei possibili obiettivi; occorrerebbe inoltre fare in modo che le installazioni militari non fossero situate nei pressi di reattori a uso civile.

A chiusura di questa analisi, si può osservare che, rispetto al più disastroso incidente ipotizzabile a un reattore nucleare, anche una sola arma nucleare può contaminare con la ricaduta radioattiva una zona molto più ampia. Non ha quindi fondamento il diffuso timore dei rischi connessi alla produzione di energia elettrica con impianti nucleari. Indubbiamente, un incidente catastrofico a un reattore causerebbe considerevole scompiglio nelle immediate vicinanze e probabilmente creerebbe problemi sanitari di lunga durata e darebbe luogo anche a qualche caso letale. Ciò nonostante, gli effetti complessivi dell'incidente possono essere modesti, in quanto anche nell'area contaminata resterebbero indenni e in piena efficienza tutti i servizi essenziali civili, sociali e sanitari. Inoltre, i rischi connessi ai reattori possono essere minimizzati attraverso un uso responsabile della relativa tecnologia. Un attacco nucleare è invece fondamentalmente diverso. Non si esagera se si afferma che mai l'umanità, in tutta la sua storia, si è trovata di fronte a un pericolo di distruzione e di morte di così vasta portata come una guerra nucleare.

L'origine dell'informazione genetica

Confermata in laboratorio, la teoria della selezione naturale delle molecole prebiotiche permette di capire come gli antichi geni a RNA interagissero con le proteine e come si sia evoluto il codice genetico

di Manfred Eigen, William Gardiner, Peter Schuster e Ruthild Winkler-Oswatitsch

Charles Darwin riconobbe nella diversità delle specie i meccanismi dell'evoluzione che operano nella generazione delle specie stesse: la variazione, la competizione e la selezione. Dai tempi di Darwin, abbiamo raggiunto un grado di comprensione della biologia molecolare, della geofisica e della geochimica dell'era prebiotica, che sarebbe stato impensabile nel diciannovesimo secolo. È possibile ora ricostruire l'evoluzione fino all'era in cui non vi erano ancora organismi veri e propri?

Una prima risposta è no. I resti fossili prebiotici, per quanto se ne sa, si sono decomposti o sono stati distrutti dalle forme di vita successive. I fossili «intellettuali» che rimangono (il codice genetico, i messaggi genetici degli organismi attuali e le catene metaboliche note) ci forniscono informazioni così frammentarie che sarebbe impossibile ricostruire l'evoluzione prebiotica così dettagliatamente come, per esempio, l'evoluzione dei primati.

Un'informazione frammentaria, tuttavia, non ha mai rappresentato una barriera insormontabile per la scoperta delle leggi della natura. Newton scoprì le leggi universali del moto dall'osservazione di pochi pianeti; Mendeleev scoprì la struttura della tavola periodica degli elementi a partire dalla conoscenza delle proprietà chimiche di pochi elementi; oggi i fisici inferiscono le leggi che descrivono le interazioni delle particelle elementari dall'osservazione di un piccolo numero di eventi. Non è necessario conoscere nei particolari la storia delle condizioni e degli eventi prebiotici per scoprire le leggi evolutive che hanno prodotto la vita sulla Terra. Si deve solo sperare di trovare documentazioni fossili superstiti sufficienti per orientare la ricerca teorica e mettere alla prova le capacità predittive delle varie teorie. In questo senso la risposta alla domanda precedente è sì: è possibile fare precise affermazioni sulle leggi naturali che hanno determinato l'origine e l'evoluzione prebiotica della vita.

In questo articolo descriviamo quanto deve essere aggiunto alle idee di Darwin per ricostruire l'evoluzione prima della comparsa degli organismi. Per prima cosa dimostreremo che i concetti darwiniani si applicano all'evoluzione a livelli molto più bassi di quello di organismo. Per spiegare l'origine della complessità degli organismi superiori e della varietà delle specie, Darwin propose che il più complesso si generasse dal meno complesso per azione della selezione naturale. Per quale motivo questo principio non dovrebbe applicarsi anche alla complessità delle macromolecole? Descriveremo le condizioni necessarie e sufficienti per il funzionamento della selezione naturale a livello molecolare. Il risultato di questa selezione è una «regolarità di eventi» che si determina inevitabilmente ogniqualvolta determinate condizioni vengono soddisfatte.

La competizione è la base della selezione naturale a livello degli organismi, ma da sola, in era prebiotica, non sarebbe stata sufficiente per selezionare le organizzazioni molecolari più adatte; certe forme di cooperazione devono essere state altrettanto essenziali. L'interazione, nell'evoluzione, fra competizione molecolare e cooperazione molecolare, rifletteva la necessità di elaborare e utilizzare la primitiva informazione genetica in modo da stabilizzarla e migliorarla. È impossibile ricreare gli stadi effettivi del miglioramento dell'informazione genetica perché durante le prime fasi dell'evoluzione furono provate e scartate un numero enorme di mutazioni casuali. Ciononostante, è possibile ora comprendere le leggi naturali che devono aver governato quegli stadi dell'evoluzione. Queste leggi possono venire messe alla prova in diversi modi: per mezzo di esperimenti con i virus dei batteri, per mezzo dello studio chimico dei componenti degli acidi nucleici e delle proteine e per mezzo dell'analisi comparativa degli acidi nucleici e delle proteine che sono sopravvissuti a tre o quattro miliardi di anni di evoluzione molecolare.

La Terra prima della comparsa della vita

Prima che il dramma della vita potesse avere inizio era necessario stabilire il palcoscenico e assicurare l'intervento di alcuni attori minori.

Lo scenario era da qualche parte sulla Terra primitiva, dove la temperatura era non troppo differente da quella attuale. La composizione della superficie terrestre era molto simile a quella attuale per quanto riguarda l'abbondanza relativa degli elementi, ma radicalmente diversa per quanto riguarda il modo in cui gli elementi erano combinati. Gli esperimenti hanno dimostrato che quasi ogni fonte di energia, come fulmini, onde d'urto, radiazioni ultraviolette o ceneri vulcaniche calde, avrebbe provocato la conversione di quantità significative di materiali superficiali primitivi in una grande varietà di sostanze che oggi si considererebbero organiche. Nel primitivo sistema solare erano presenti quantità molto ingenti di materiali provenienti da comete e meteoriti, che possono aver contribuito in modo sostanziale alla formazione della superficie della Terra. Gli effetti della radiazione solare su questo materiale estremamente freddo, rimasto dalla condensazione del sistema solare, potrebbero aver prodotto molecole organiche di dimensioni pari a quelle di alcuni polimeri biologici.

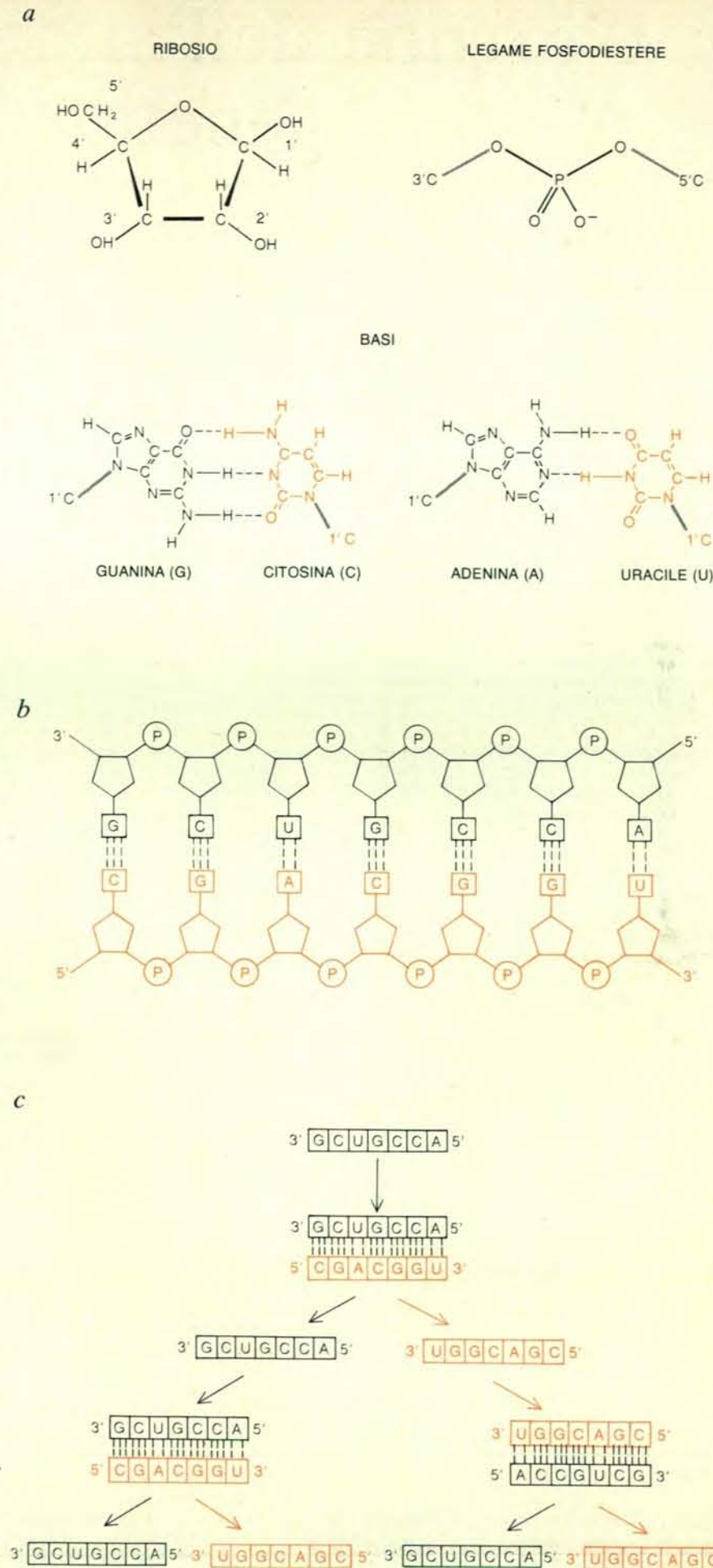
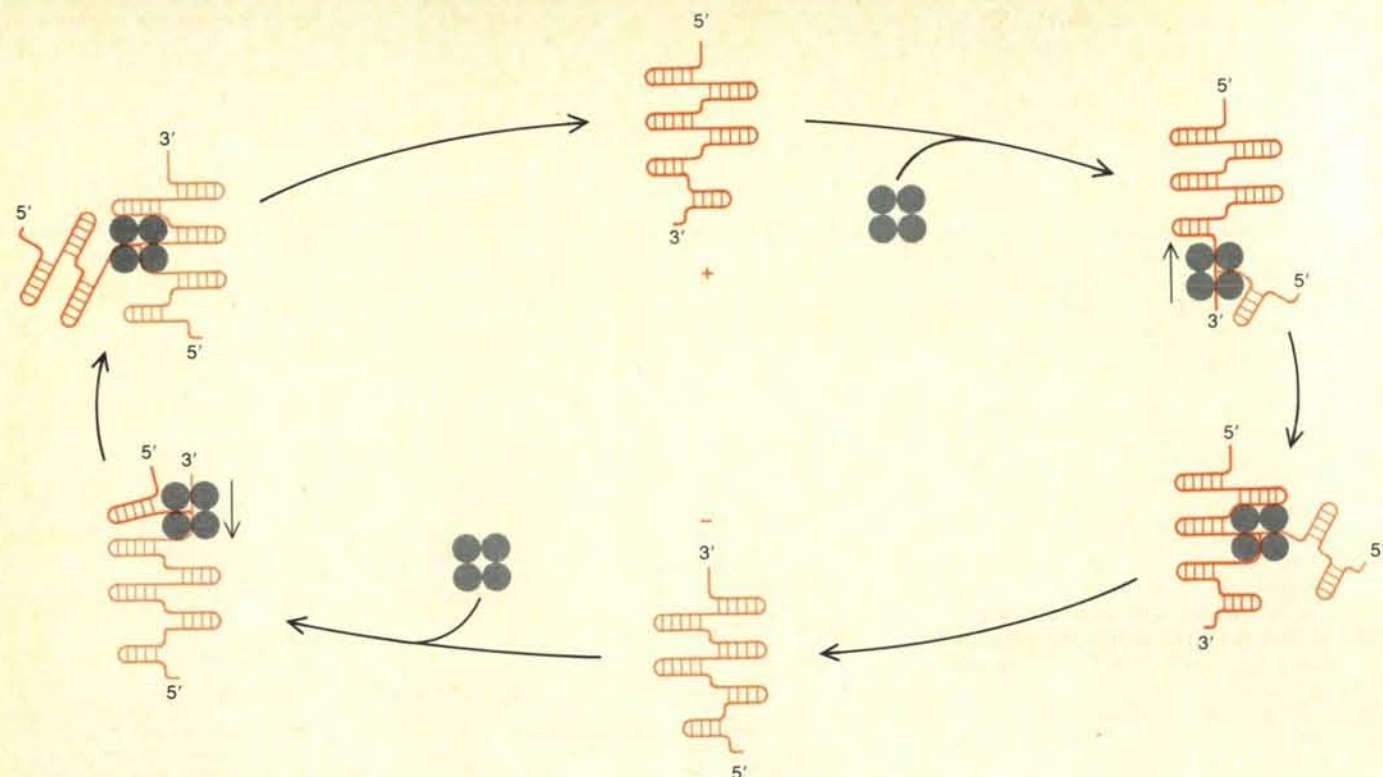
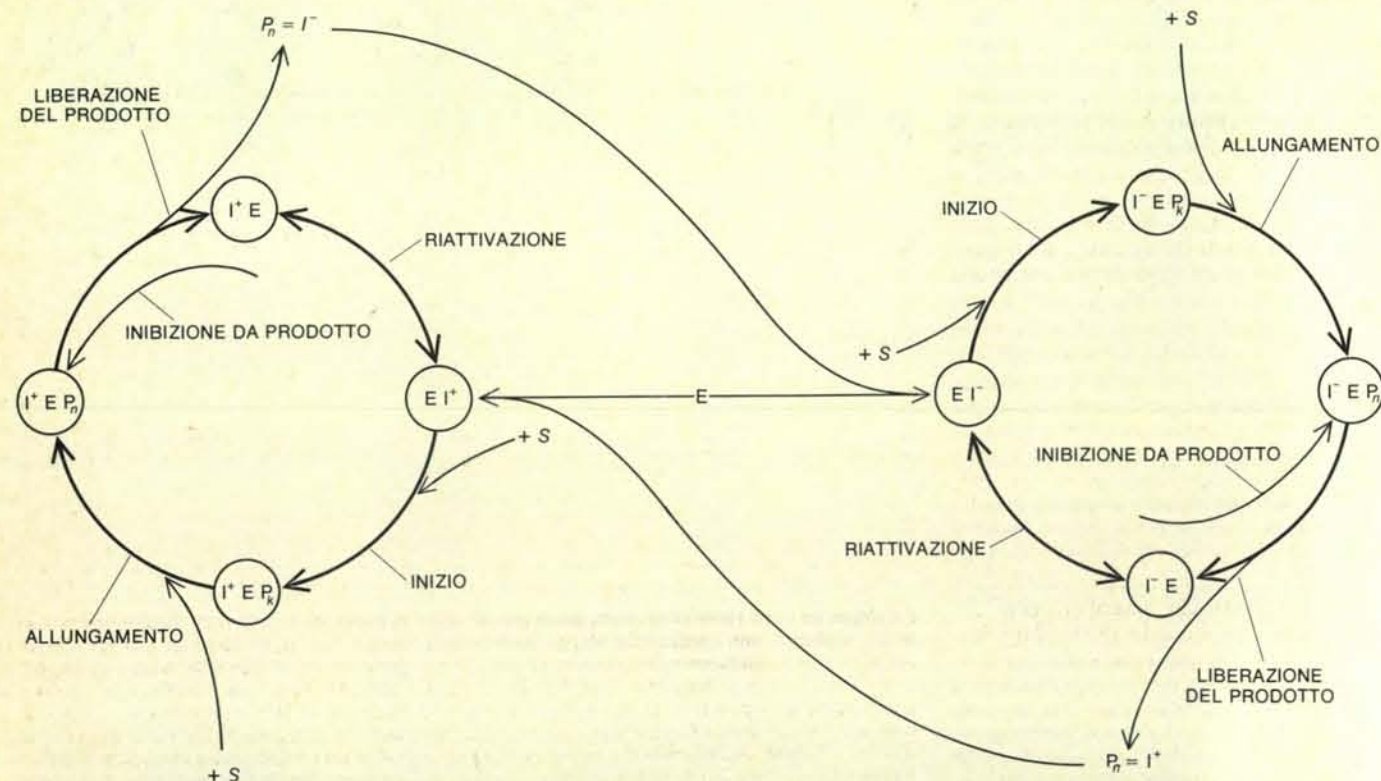
Il «brodo primordiale» da cui ha avuto origine la vita includeva non solo quegli amminoacidi, quegli zuccheri e quelle sostanze particolari che sono oggi i reagenti biochimici essenziali, ma anche molte altre molecole che oggi sono solo curiosità di laboratorio. Era perciò necessario che il primitivo «principio organizzatore» possedesse una forte selettività fin dall'inizio. Era necessario tollerare un sovraccarico enorme di piccole molecole «sbagliate» biologicamente, ma possibili chimicamente. Da questo brodo il principio organizzatore doveva essere in grado di estrarre quelle molecole che alla fine sarebbero diventate i monomeri standard dei polimeri biologici, e doveva poterli legare affidabilmente in configurazioni particolari.

La quantità totale di materiale organico potenziale era immensa. Se il carbonio che oggi si trova nel carbone, nelle rocce carbonatiche e nel materiale vivente fosse stato uniformemente distribuito nell'acqua degli oceani, la soluzione di carbonio risultante sarebbe stata concentrata come un buon brodo di carne. I processi geofisici come l'erosione, l'evaporazione e la sedimentazione devono aver agito allora come oggi e aver creato una varietà di ambienti. Evidentemente, almeno uno di questi ambienti era adatto, per temperatura e composizione chimica, all'origine della vita.

Il brodo primordiale dovette affrontare una crisi energetica: le primitive forme di vita avevano bisogno di estrarre in qualche modo l'energia chimica dalle molecole che componevano il brodo. Per la nostra storia non è importante conoscere il modo in cui lo fecero; è possibile ammettere l'esistenza di qualche sistema di conservazione e utilizzazione dell'energia basato sui fosfati. La rigenerazione non metabolica della riserva di energia del sistema dei fosfati (forse per mezzo di qualche tipo di conversione dell'energia solare in energia chimica) dovette durare fino all'evoluzione di un meccanismo per la fermentazione di qualche componente non indispensabile del brodo. La fermentazione fu probabilmente sufficiente finché l'avvento della fotosintesi non fornì una fonte rinnovabile e continua di energia.

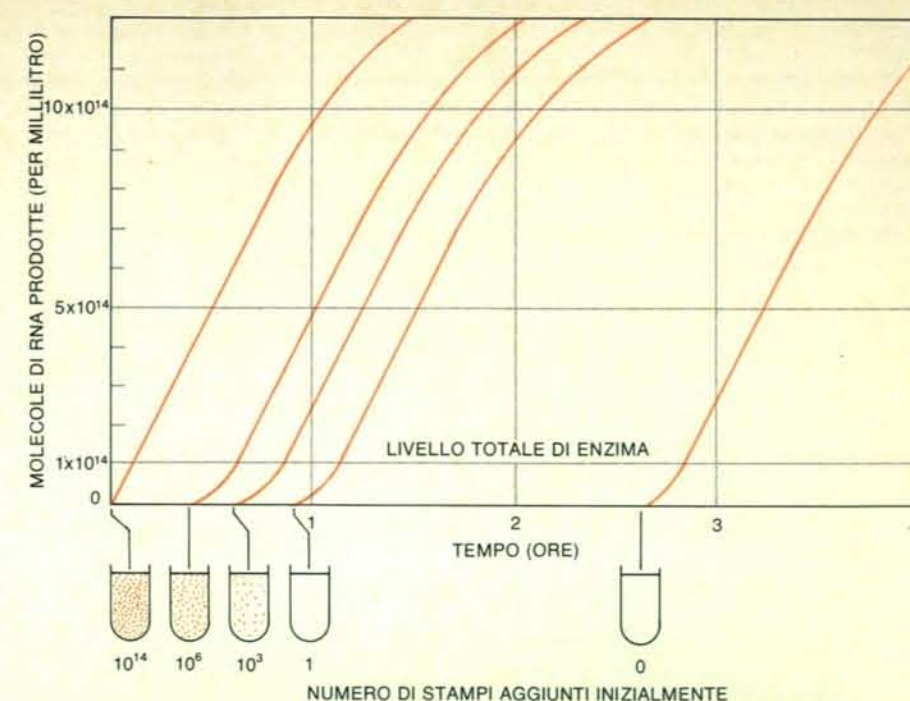
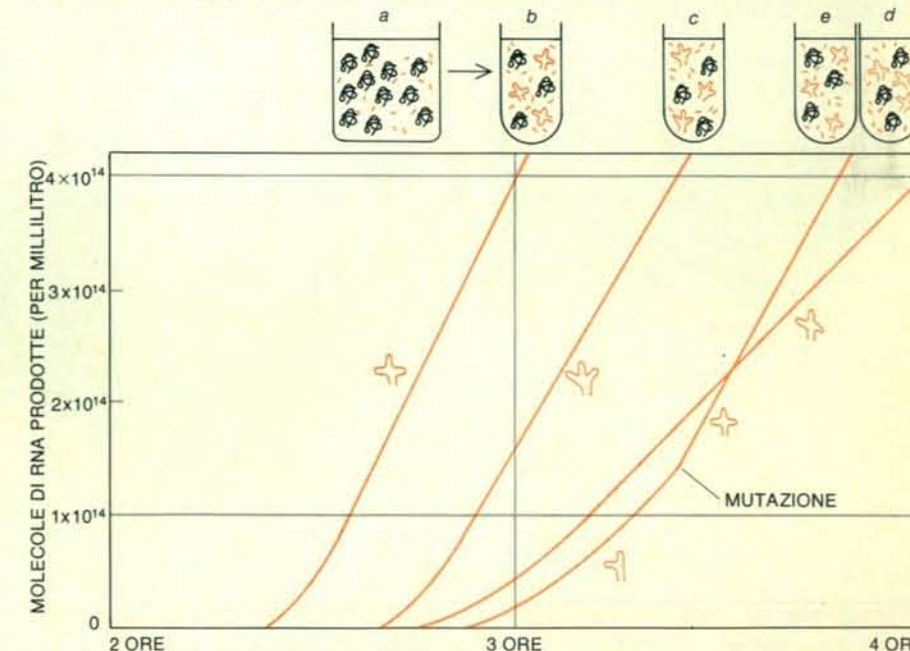
I primi geni

Nelle cellule l'informazione genetica è conservata nel DNA, trascritta nell'RNA messaggero e successivamente tradotta in proteine; nei virus l'informazione è conservata su filamenti di DNA oppure di RNA. Entrambi gli acidi nucleici sono molecole molto lunghe composte di nucleotidi. Ogni nucleotide ha tre componenti: un gruppo chimico chiamato base, uno zucchero (il desossiribosio nel DNA, il ribosio nell'RNA) e un gruppo fosfato. Il legame alternato di zuccheri e gruppi

I primi portatori dell'informazione genetica furono le molecole di RNA: catene lineari di nucleotidi uniti da legami fosfodiestere. Ogni nucleotide dell'RNA consiste di uno zucchero (ribosio) legato ad una delle quattro basi *(a)*; l'informazione è codificata dalla particolare sequenza di basi che formano la catena di RNA. Le basi sono complementari: per mezzo di legami idrogeno *(linee tratteggiate)* l'adenina si appaia specificamente con l'uracile e la guanina si appaia con la citosina. L'informazione codificata dall'RNA viene perpetuata tramite la sintesi di un filamento complementare *(in colore)*, che costituisce la replica esatta del filamento originale che funge da stampo *(b)*. Nel disegno è raffigurato il trasferimento dell'informazione di una sequenza di RNA di sette nucleotidi per due generazioni successive *(c)*.

L'RNA a filamento singolo del virus batterico Q_β viene replicato da un enzima costituito di quattro subunità. Il filamento «+» dell'RNA virale viene normalmente mantenuto in una configurazione ripiegata specifica dall'appaiamento delle basi complementari lungo il filamento *(in alto al centro)*. Esso viene svolto dall'enzima replicativo mentre questo si muove dall'estremità 3' verso l'estremità 5' *(in alto a destra)*. L'enzima lega i monomeri (precursori dei nucleotidi in forma attivata) allo stampo seguendo le regole di appaiamento delle basi, e costruisce un filamento «—» complementare *(in basso al centro)*. Il nuovo filamento si ripiega impedendo la formazione di una doppia elica, che bloccherebbe la replicazione. A sua volta il filamento «—» viene replicato e dà origine ad un duplicato del filamento «+» originale *(a sinistra)*.

I cicli interconnessi di sintesi dei filamenti «+» e «—» sono caratteristici della replicazione dell'RNA. Il processo avviene in quattro fasi: inizio della replicazione, allungamento del filamento che viene replicato, rilascio del filamento replicato e riattivazione. E è l'enzima replicativo, I è la molecola informazionale, cioè lo stampo di RNA, P_n è il prodotto, cioè la molecola di RNA replicato, S indica il substrato, l'insieme dei monomeri nucleotidici. Le proprietà catalitiche degli elementi del sistema determinano la concentrazione relativa dei filamenti «+» e «—» in modo tale che la velocità di crescita dell'insieme dei filamenti è la media geometrica delle velocità di crescita dei singoli filamenti.

fosfato forma lo «scheletro» della molecola; l'informazione genetica è codificata da particolari sequenze di basi. Nel DNA le quattro basi sono le purine adenina *(A)* e guanina *(G)* e le pirimidine timina *(T)* e citosina *(C)*; nell'RNA l'uracile *(U)* sostituisce la timina. Le basi sono complementari e si appaiano secondo regole prestabilite: *A* si appaia con *T* (o *U*) e *G* si appaia con *C*. La complementarità è la base della replicazione e della trascrizione. Nella replicazione un filamento di DNA o RNA serve da stampo su cui i nucleotidi complementari vengono assemblati dai vari enzimi noti come polimerasi e replicasi secondo le regole di complementarità e formano un filamento complementare contenente una copia duplicata dell'informazione. Nella trascrizione una sequenza di DNA dà origine con un processo simile a un filamento complementare di RNA messaggero.

Conoscendo le proprietà chimiche di DNA e RNA, che cosa si può dedurre sull'identità dei primi portatori prebiotici dell'informazione genetica? I nucleosidi contenenti desossiribosio, di cui è costituito il DNA, sono più difficili da maneggiare chimicamente dei loro analoghi, contenenti ribosio, dell'RNA. In effetti, nelle cellule la sintesi dei monomeri del DNA procede attraverso intermedi ribonucleotidici, e la stessa replicazione del DNA comincia con la sintesi di blocchi *(primer)* di RNA. Negli organismi attuali l'informazione genetica viene decodificata da meccanismi complessi di proteine ed RNA. Perché questo meccanismo si potesse evolvere, i portatori stessi dell'informazione genetica dovevano possedere caratteristiche strutturali che li rendessero riconoscibili dalla macchina traduttrice. L'RNA a singolo filamento può ripiegarsi su se stesso e formare una grande varietà di strutture tridimensionali, in contrasto con l'uniformità della doppia elica del DNA. Nei meccanismi cellulari attuali, quando sono richieste proprietà sia funzionali sia informazionali, viene utilizzato l'RNA. Non c'è ragione di pensare che nei primi stadi della formazione della vita le cose andassero diversamente; e nulla fa pensare che esistesse un processo per mezzo del quale l'informazione conservata sotto qualunque altra forma potesse essere trasferita sugli acidi nucleici.

La ricerca della probabile identità chimica dei primi geni ci porta quindi rapidamente alle sequenze di basi dell'RNA. Si può ragionevolmente prevedere che le vie primordiali di sintesi e differenziazione producessero concentrazioni estremamente piccole di sequenze nucleotidiche che sarebbero considerate «corrette» secondo gli standard della biochimica degli organismi attuali: le sequenze primordiali avevano le stesse basi, gli stessi legami covalenti e la stessa stereochimica, cioè la stessa distribuzione nello spazio dei gruppi chimici. Queste sequenze erano presenti, tuttavia, insieme a miriadi di altre che oggi sarebbero considerate «sbagliate», con diversa stereochimica, legami scorretti e basi non standard. Che cosa c'era di così partico-

La velocità di moltiplicazione dei filamenti di RNA viene studiata incubando una miscela di filamenti di RNA stampo di Q_β, monomeri nucleotidici e Q_β replicasi. Quando lo stampo è in eccesso rispetto all'enzima, la crescita delle molecole di stampo è lineare (finché non arriva a saturazione ad alta concentrazione a causa dell'inibizione da prodotto). Questo tipo di crescita è caratteristico di replicazioni successive in cui l'enzima deve essere riattivato ogni volta. Quando esso è in eccesso rispetto allo stampo, la crescita è esponenziale; appena un ciclo di replicazione è terminato, i filamenti di prodotto neosintetizzato e il vecchio filamento di RNA stampo si legano a una molecola libera di enzima, dando origine a due repliche. Riducendo la concentrazione di stampo, la curva di crescita si sposta a destra. Anche se non è presente inizialmente alcuno stampo, un RNA correlato ai frammenti di Q_β viene sintetizzato *ex novo* dopo un lungo periodo di incubazione a causa delle interazioni tra enzima e monomeri.

***La sintesi ex novo* viene dimostrata da un esperimento di clonazione. Una miscela di monomeri e di Q_β replicasi, non contenente alcuna molecola di stampo *(a)*, viene incubata per un tempo sufficiente ad amplificare ogni stampo che possa essere presente come impurità, ma per un tempo troppo breve per permettere la sintesi *ex novo*. La miscela viene poi suddivisa in quattro compartimenti separati *(b-e)*, in ognuno dei quali i filamenti di RNA sintetizzati *ex novo* si moltiplicano a velocità diverse *(curve)*. In ogni compartimento la selezione naturale genera uno stampo ottimale, che alla fine diventa il solo prodotto presente; tuttavia, dato che la divisione in compartimenti ha preceduto la selezione, i quattro compartimenti alla fine contengono prodotti diversi. Talvolta una mutazione avviene abbastanza tardi da poter essere osservata, come nel compartimento *e*. In contrasto con la natura altamente riproducibile della replicazione diretta dallo stampo, nel tempo richiesto per la comparsa dei prodotti sintetizzati *ex novo* si osservano ampie fluttuazioni. Queste riflettono il fatto che solo un piccolo numero di molecole è coinvolto nel passaggio che condiziona la velocità della reazione complessiva: la sintesi del primo stampo.**

lare nelle sequenze che assomigliavano all'RNA attuale?

La risposta è semplice: i filamenti di RNA con una stereochimica omogenea e una struttura corretta di legami nello scheletro della molecola potevano formare riproducibilmente strutture secondarie stabili; la molecola si ripiegava sempre nello stesso modo a causa della formazione di ponti idrogeno tra coppie di nucleotidi complementari. Questo era un vantaggio importante, perché rendeva i filamenti di RNA più resistenti all'idrolisi, la rottura a opera di una molecola d'acqua che è il destino ultimo di ogni polimero in soluzione acquosa.

I filamenti primordiali di RNA che fortuitamente possedevano le scheletro giusto e i nucleotidi giusti avevano un secondo vantaggio cruciale: erano i soli capaci di autoreplicarsi stabilmente. Essi erano contemporaneamente la fonte dell'informazione (attraverso le regole di appaiamento) e l'oggetto fisico da replicare secondo le istruzioni. Qui, a livello molecolare, sta la soluzione del vecchio dilemma dell'uovo e la gallina. Cosa venne per prima, la funzione o l'informazione? Come dimostreremo, nessuna delle due può aver preceduto l'altra; esse dovettero evolversi insieme.

Le specie chimiche e le reazioni dei tempi prebiotici sicuramente avevano un certo numero di caratteristiche in comune con le reazioni biochimiche moderne. Sidney Fox e i suoi colleghi all'Università di Miami hanno dimostrato, per esempio, che alcune funzioni enzimatiche possono venire svolte da polimeri «proteinoidi» sintetizzati semplicemente scaldando una miscela di amminoacidi (i costituenti delle proteine). Oltre a tali primitivi catalizzatori vi erano sicuramente molecole reattive alla stimolazione con luce solare; vi erano lipidi (grassi) o molecole simili in grado di formare strutture membranose e vi erano forse persino polisaccaridi, polimeri di zuccheri che costituivano potenziali risorse di energia. In breve, per mezzo di sequenze di reazioni non biologiche, era stata creata una enorme varietà di molecole funzionali.

Queste molecole funzionali possono essere state importanti nella chimica del brodo primordiale. Tuttavia, non potevano evolversi. La loro efficienza accidentale si fondava su restrizioni strutturali non accidentali, come interazioni favorevoli con le molecole circostanti o particolari strutture tridimensionali. Se dovevano essere in grado di migliorare la loro efficienza, e se varianti più funzionali dovevano essere selezionate a partire da varianti meno funzionali, le molecole dovevano superare queste restrizioni strutturali. Solo le molecole in grado di autoreplicarsi e di conservare l'informazione potevano farlo. Discuteremo ora come il contenuto informazionale di queste molecole possa migliorare e la loro complessità aumentare e come esse soppiantino le varianti meno funzionali.

L'autoreplicazione

Il virus Q_β, che infetta il batterio *Escherichia coli*, costituisce un modello sperimentale per gli studi sull'autoreplicazione delle macromolecole. Possiede un genoma (il corredo completo dell'informazione genetica) costituito da un singolo filamento di RNA lungo circa 4500 nucleotidi. Solo parte della molecola costituisce il messaggio genetico; il resto svolge una varietà di funzioni strutturali (anziché informazionali); per esempio, costituisce il bersaglio per il riconoscimento specifico da parte di particolari enzimi. Alcuni anni fa Sol Spiegelman, che allora si trovava all'Università dell'Illinois, isolò l'enzima replicativo di Q_β e dimostrò che era in grado di riprodurre l'RNA virale in un sistema *in vitro* e generare copie infettive. Isolò anche, dalle cellule di *E.coli* infettate dal virus, una molecola di RNA «satellite» lunga 220 nucleotidi, che viene replicata dalla Q_β replicasi con efficienza estremamente alta. L'RNA satellite e altre «minivarianti» simili costituiscono, insieme alla Q_β replicasi, un comodo modello sperimentale per lo studio della replicazione dell'RNA.

Un esperimento tipico viene realizzato con una soluzione contenente ioni magnesio, una piccola concentrazione di Q_β replicasi altamente purificata e i quattro monomeri precursori dell'RNA in forma attivata. La forma attivata è rappresentata dai nucleosidi trifosfati, *ATP, GTP, UTP*, e *CTP*, in cui la base e lo zucchero sono legati a una «coda» di tre unità fosfato. Uno dei quattro nucleosidi trifosfati (di solito il *GTP*) è marcato con un isotopo radioattivo per poter misurare la sintesi del nuovo RNA. Per iniziare la replicazione si aggiunge una certa quantità di RNA stampo e si incuba la miscela.

In alcuni esperimenti di replicazione eseguiti nel 1974 da Manfred Sumper nel nostro laboratorio al Max Planck Institut für Biophysische Chemie a Gottinga avvenne qualcosa di assolutamente inaspettato. Quando Sumper iniziava l'incubazione con più stampo che enzima, trovava un aumento lineare della concentrazione di RNA, che continuava fino al raggiungimento di una concentrazione relativamente alta. Questo indicava che tutte le molecole di enzima potevano venire occupate simultaneamente dagli stampi che venivano replicati; malgrado la concentrazione di stampi crescesse costantemente, la concentrazione dei complessi enzima-stampo rimaneva costante. Per questo motivo il nuovo RNA continuava a venire sintetizzato a velocità costante.

Il passo successivo fu ovviamente quello di ridurre la quantità dell'RNA stampo nella miscela iniziale. Questo provocò lo spostamento della curva di crescita lineare verso tempi maggiori, senza cambiare la forma e la pendenza della curva. Riduzioni successive ritardarono la crescita di un fattore proporzionale al logaritmo della concentrazione dello stampo. In altre parole, la diluizione da 10^6 a 10^4 molecole di stampo per provetta spostava la curva di un fattore uguale alla diluizione da 10^4 a 10^2 molecole. Questa relazione logaritmica indicava chiaramente che quando l'enzima è in eccesso rispetto allo stampo, ogni molecola di RNA appena sintetizzato va a occupare immediatamente una molecola libera di enzima. La concentrazione dello stampo aumenta così esponenzialmente, invece che linearmente, in funzione del tempo. La reazione funziona anche con microscopiche quantità iniziali di stampo; in effetti avviene anche in presenza di una singola molecola di stampo per provetta. (Un metodo per la clonazione di singole molecole è basato su questa scoperta.)

Immaginate la nostra sorpresa quando Sumper un bel giorno riferì che anche senza l'aggiunta iniziale di una molecola di RNA stampo veniva sintetizzato del nuovo RNA, sebbene dopo tempi di incubazione molto più lunghi e variabili. La possibilità che una delle molecole di enzima fosse inquinata da qualche impurità di RNA fu rapidamente scartata per mezzo di vari esperimenti. I nucleosidi trifosfati che fungevano da substrato furono sottoposti a condizioni in cui qualsiasi polimero sarebbe stato completamente degradato. Gli enzimi furono purificati e analizzati con la massima cura possibile. Aggiungemmo deliberatamente singole molecole di stampo come impurità e dimostrammo che queste causavano un tipo di crescita completamente differente. Alla fine ci convincemmo che di fronte a noi stavano molecole di RNA che erano state sintetizzate *ex novo* dalla Q_β replicasi. Per di più, il prodotto sintetizzato *ex novo* aveva una composizione costante, che in ogni prova si dimostrò essere simile o persino identica alla minivariante di Spiegelman.

Gli studi comparati delle velocità di reazione dimostrarono ben presto che i meccanismi della sintesi indotta dallo stampo e di quella *ex novo* erano completamente differenti. Il complesso meccanismo della sintesi dipendente dallo stampo fu risolto in reazioni parziali per mezzo delle quali i risultati dei nostri esperimenti sulle velo-

INIZIO
PRIMO TRASFERIMENTO
SECONDO TRASFERIMENTO
ULTIMO TRASFERIMENTO
INCUBAZIONE PER IL TEMPO T_1
INCUBAZIONE PER IL TEMPO T_2
INCUBAZIONE PER IL TEMPO T_n

Il metodo dei trasferimenti successivi, ideato da Sol Spiegelman del Columbia University College of Physicians and Surgeons per prolungare la crescita indefinitamente, è stato applicato per dimostrare l'inizio della sintesi *ex novo* dell'RNA e l'evoluzione degli stampi ottimali. Una serie di provette viene predisposta con la Q_β replicasi, i fattori di crescita richiesti e i monomeri *A, U, G, C*, ma nessuno stampo di RNA. La miscela contenuta nella prima provetta viene incubata alzando la temperatura; dopo un lungo tempo di incubazione viene sintetizzata una collezione eterogenea di brevi stampi. Una piccola frazione di questa miscela viene trasferita nella seconda provetta e incubata. La procedura viene ripetuta numerose volte; alla fine, viene selezionato un unico stampo ottimale. Per rendersi conto della enorme moltiplicazione degli stampi resa possibile da questo metodo, si consideri che, se la crescita in ognuna delle provette della serie amplifica 10 000 volte l'RNA prodotto, 10 trasferimenti sono equivalenti a una crescita che sarebbe sufficiente a saturare di RNA tutti gli oceani del globo.

Il modello della quasispecie

La chimica prebiotica fornì l'ambiente adatto per l'evoluzione darwiniana: popolazioni di molecole autoreplicanti di RNA entrarono in competizione per una disponibilità limitata di «cibo» (monomeri ricchi di energia). La generazione continua di sequenze mutanti, alcune delle quali dotate di proprietà vantaggiose, impose la selezione delle molecole evolutivamente migliori. Questa competizione darwiniana a livello molecolare è descrivibile con un modello matematico.

Sia N_i il numero di nucleotidi che compongono la sequenza i; la posizione di ogni nucleotide è identificata da un indice p che può assumere ogni valore compreso tra 1 e N_i. Sia q_{ip} la probabilità che il nucleotide nella posizione p della sequenza i venga copiato correttamente nel processo di autoreplicazione; la frequenza di errore per quella posizione è allora $(1\text{-}q_{ip})$. Il simbolo q_{ip} descrive in questo modo la qualità, o meglio la fedeltà, della replicazione alla posizione p della sequenza i. La probabilità Q_i che una replicazione generi una sequenza i corretta è data dal prodotto di tutte le fedeltà replicative di ogni singolo nucleotide:

$$Q_i = q_{i1} \times q_{i2} \cdots \times q_{iN_i} = \overline{q}_i^{N_i}$$

dove $\overline{q}_i$ è la media geometrica delle fedeltà replicative della sequenza i.

La sequenza i può sopravvivere a replicazioni successive solo se gli errori di copiatura non si accumulano. Questo richiede che la sequenza superi, per quanto riguarda la crescita netta, la media dei suoi competitori per un fattore di vantaggio competitivo S_i. Inoltre, i può venire selezionata solo se è soddisfatta una condizione di sopravvivenza chiamata «limite di errore»; esso è rappresentato da $Q_i S_i$ e deve essere maggiore di 1 per permettere la sopravvivenza della sequenza i.

La crescita netta è descritta da un'equazione che rappresenta come x_i (la frazione di tutte le sequenze presenti che sono copie esatte della sequenza i) vari in funzione del tempo. Le cause principali della variazione di x_i sono la replicazione corretta di i e la replicazione errata di sequenze strettamente correlate, designate collettivamente come j, che possono dare origine ad i per mutazione. La velocità di variazione di x_i è data da:

$$(W_{ii} - E)\, x_i + \text{somma delle } W_{ij}\, x_j$$

In questa espressione W_{ii} rappresenta la velocità di replicazione corretta della sequenza i ed $\overline{E}$ la velocità media di produzione netta (differenza tra le sequenze replicate e quelle perdute) di tutte le sequenze presenti; entrambe le velocità sono espresse come velocità per copia. W_{ij} è la velocità di produzione della sequenza i a causa di errori nella replicazione di j, espressa come velocità per copia di j. I contributi di tutte le sequenze designate come j sono sommati. Quindi il primo termine rappresenta la velocità di competizione della sequenza i con altre sequenze e il secondo la velocità di produzione di i per mezzo di mutazioni di altre sequenze.

Queste velocità determinano il modo in cui ogni insieme arbitrario di sequenze si organizza. Il primo termine può essere positivo o negativo a seconda che W_{ii} sia maggiore o minore della produzione netta media E. Se W_{ii} è maggiore, x_i cresce; se è minore, x_i decresce costantemente finché la sequenza i si estingue o viene prodotta solo per mutazione. Il declino delle sequenze per le quali W_{ii} è minore di E, tuttavia, ha l'effetto di aumentare la velocità di produzione netta media E; questo rende ancora più difficile per le sequenze sopravvissute soddisfare la condizione che W_{ii} sia maggiore di E, e quindi crescere. La competizione autoreplicativa assomiglia a una competizione di salto in alto in cui l'asticella viene alzata finché rimane in gara un solo competitore. Nella competizione molecolare, tuttavia, non emerge mai un unico vincitore: a causa dei termini «mutazionali» $W_{ij}\, x_j$, il competitore più forte produce costantemente sequenze mutanti con le quali deve continuare a competere. All'equilibrio, il competitore migliore, designato «sequenza principale» m, coesiste con tutte le sequenze mutanti derivate da lui stesso a causa di replicazioni erronee. Noi definiamo «quasispecie» questa distribuzione di sequenze.

Questa analisi dimostra che il principio darwiniano di selezione naturale non è di natura assiomatica, ma deriva dall'insieme di condizioni fisiche a cui è soggetta l'autoreplicazione. Il risultato finale della selezione, la quasispecie, è stabile finché una nuova mutazione non produce casualmente una sequenza con velocità di crescita superiore a quella della sequenza principale esistente (o finché un cambiamento delle condizioni ambientali non produce un effetto equivalente). Quando questo avviene, la nuova sequenza di maggior successo prolifera finché (insieme ai suoi mutanti) arriva a prendere il posto della vecchia quasispecie, che scompare.

Sono stati ricavati i parametri che descrivono la struttura della quasispecie di RNA. Per esempio, la lunghezza massima della sequenza principale, in nucleotidi, è:

$$\frac{2{,}3 \log S_m}{1 - \overline{q}_m}.$$

Sequenze principali più lunghe non sono in grado di superare il limite di errore, cioè il loro $Q_m S_m$ non è superiore a 1.

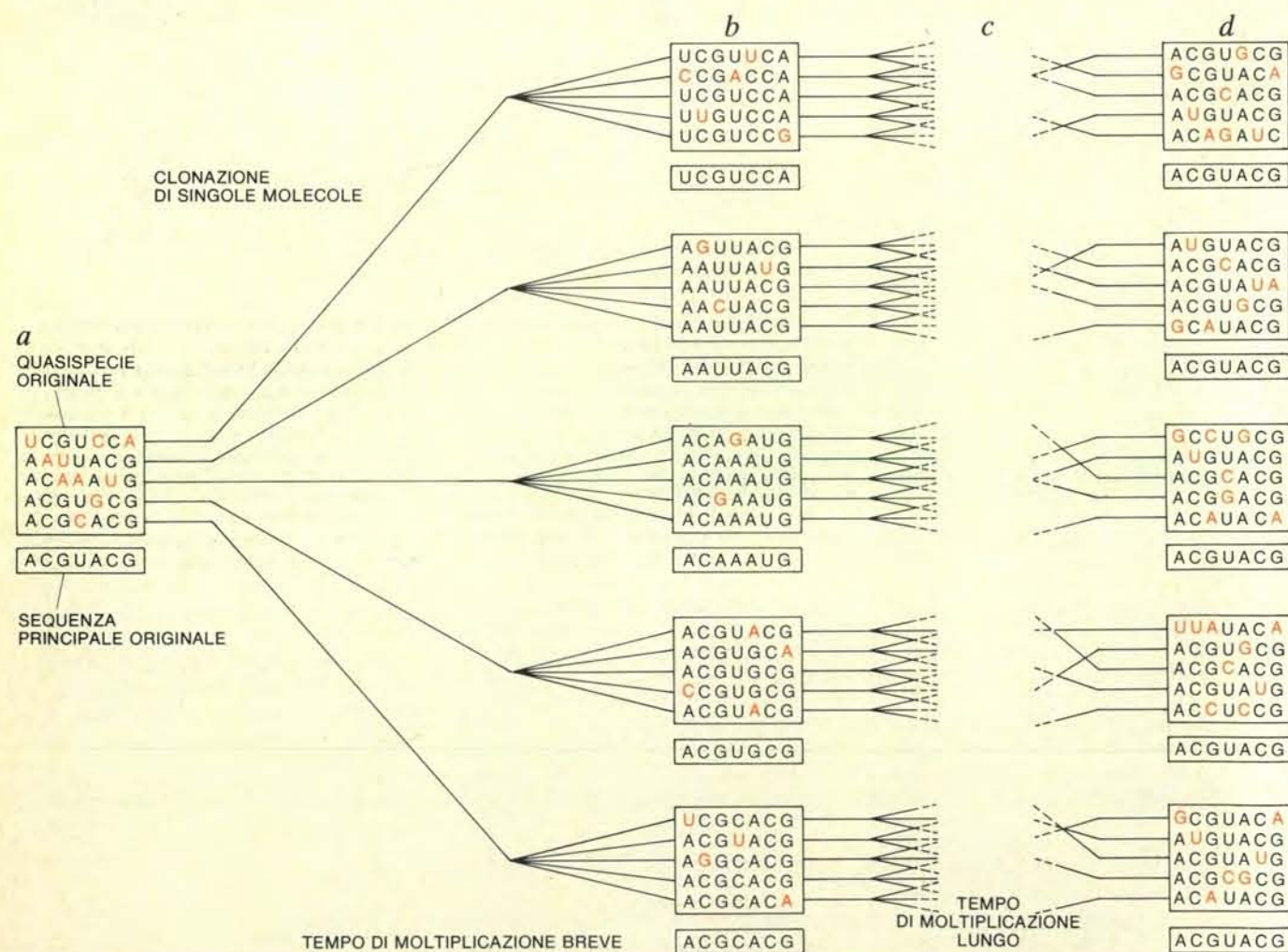
Questa discussione riassume i risultati principali delle ricerche matematiche condotte dal nostro gruppo e in seguito da B. L. Jones, R. H. Enns e S. S. Ragnekar della Simon Fraser University nella Columbia Britannica e da C. J. Thompson e J. L. McBride dell'Università di Melbourne.

cità di reazione poterono essere rappresentati algebricamente per mezzo di equazioni. Nella reazione dipendente dallo stampo una molecola di enzima è associata a una molecola di stampo e i monomeri substrato vengono aggiunti uno alla volta durante l'allungamento della copia; non si osserva alcun tipo di cooperazione tra i monomeri substrato. La sintesi *ex novo*, al contrario, richiede la cooperazione di almeno tre o quattro monomeri substrato nella reazione parziale più lenta. Inoltre almeno due molecole di enzima, ciascuna caricata con il substrato, partecipano a questa reazione parziale. Una delle molecole di enzima, evidentemente, sostituisce la molecola di stampo che manca, esponendo in maniera corretta i substrati legati all'enzima polimerizzante.

Spiegelman e Donald R. Mills del Columbia University College of Physicians and Surgeons hanno determinato la sequenza completa della minivariante da 220 nucleotidi. Analizzando la sequenza ci accorgemmo che poteva essere rappresentata come composta (tranne 56 mutazioni e 2 inserzioni) da copie multiple di quattro tetrameri e due trimeri. I tetrameri sono *CCCC* e *UUCG* e i loro complementari, *GGGG* e *CGAA*; i trimeri, *CCC* e il loro complementare *GGG*, rappresentano forme troncate dei tetrameri. La sequenza *CCC* è stata identificata da Sumper e Bernd-Olaf Küppers come la sequenza di riconoscimento che deve essere presente in tutti gli RNA che interagiscono specificamente con la Q_β replicasi; *UUCG* è una sequenza di basi che, in un contesto diverso (la traduzione dell'RNA messaggero in proteine), si lega a una delle proteine che funzionano da subunità del complesso della Q_β replicasi.

La scoperta della sintesi di RNA *ex novo* viola forse il dogma centrale della biologia molecolare, secondo il quale l'informazione viene trasferita unicamente dagli acidi nucleici alle proteine, e non viceversa? La selezione di questi particolari trimeri e tetrameri rappresenta chiaramente una «istruzione» impartita dalle proteine del complesso della Q_β replicasi. Tuttavia gli stessi trimeri e tetrameri avrebbero potuto dare origine a un'enorme varietà di altre sequenze oltre a quella che noi effettivamente ritrovavamo al termine della reazione, amplificata fino a 10^{12} volte. Non era forse possibile che tutto ciò fosse provocato dalla selezione naturale, piuttosto che da istruzioni provenienti dalle proteine?

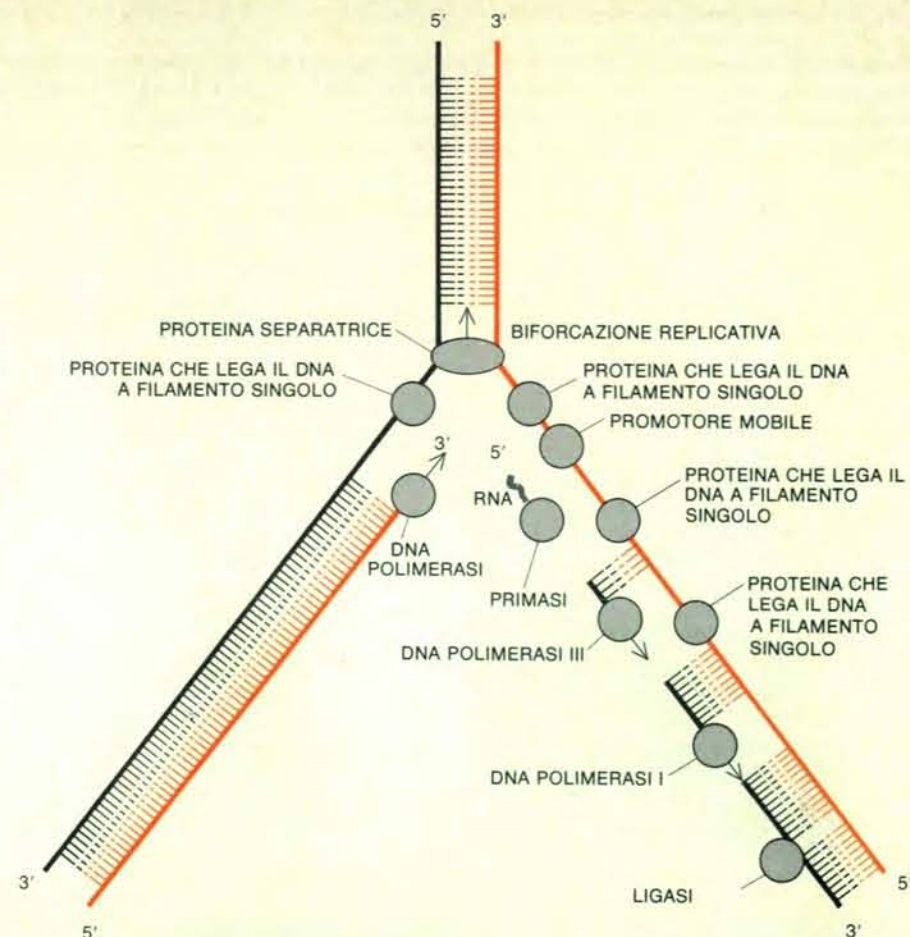
Il ruolo della selezione

Recentemente Christof Biebricher e Rüdiger Luce hanno realizzato, nel nostro laboratorio, un esperimento decisivo che ha fornito una risposta alla domanda. La loro impostazione è stata determinata

La sequenza principale di una quasispecie è conservata malgrado la continua ridistribuzione delle sequenze mutanti, come è stato dimostrato da Charles Weissmann dell'Università di Zurigo. Weissmann ha clonato l'RNA di Q_β nei batteri, diluendo una quasispecie originale del virus *(a)* così fortemente che ogni infezione era originata da una singola particella virale. Successivamente ha analizzato la sequenza di RNA di ciascun clone *(b)* per mezzo del *«finger printing»* (letteralmente, registrazione delle impronte digitali), una elettroforesi bidimensionale dell'RNA parzialmente frammentato. Ogni clone è stato poi fatto moltiplicare per molte generazioni *(c)* e sottoposto in questo modo a una forte pressione selettiva. Il risultato è stato lo stabilirsi di nuove distribuzioni quasispecifiche *(d)*. Il principio dell'esperimento è illustrato qui per cinque sequenze originali, ognuna lunga sette nucleotidi. La sequenza principale è quella in cui ogni posizione è occupata dal nucleotide più frequente in quella data posizione nella distribuzione quasispecifica. Le differenze dalla sequenza principale sono indicate in colore. La sequenza principale è sempre univocamente definita anche se è presente con frequenza molto bassa tra le sequenze mutate. Dopo la clonazione le nuove sequenze principali *(b)* sono tutte diverse tra loro, e nessuna corrisponde alla sequenza principale originale. Dopo un numero estremamente elevato di generazioni, tuttavia, tutte le sequenze principali sono riconvertite nella sequenza originale *(d)*.

dalla particolare cinetica della sintesi *ex novo*. Essi incubarono una miscela di reazione senza stampi, per un tempo abbastanza lungo da amplificare ogni impurità presente che potesse svolgere il ruolo dello stampo, ma troppo breve per la formazione dell'RNA sintetizzato *ex novo*. La miscela veniva poi divisa in compartimenti separati in cui venivano mantenute le condizioni ideali per la sintesi *ex novo*. Il risultato fu chiaro: sebbene ogni compartimento contenesse una popolazione uniforme di prodotti sintetizzati *ex novo*, questi differivano da compartimento a compartimento. Le analisi successive rivelarono comunque che le sequenze non mancavano del tutto di relazioni reciproche.

Quando i compartimenti venivano incubati separatamente, si osservavano ampie variazioni nei tempi di comparsa delle diverse popolazioni. Le fluttuazioni riflettono la natura probabilistica di un processo in cui la sintesi di una singola molecola è il passaggio che determina la velocità della reazione globale. Al contrario, l'amplificazione dello stampo procede sempre deterministicamente, con costanti cinetiche ben definite, anche se la reazione comincia con una sola o poche molecole di stampo. Le piccole fluttuazioni nella velocità di amplificazione tendono a compensarsi durante i cicli ripetuti di replicazione.

I primi prodotti che comparivano nei diversi compartimenti chiaramente non erano ancora ottimizzati da alcun processo evolutivo. Alcuni erano corti, fino a un minimo di 60 nucleotidi, e prodotti ancora più corti prevalevano forse nei primi stadi dell'amplificazione. (Una quantità analizzabile di RNA è rappresentata da almeno 10^{12} molecole. Questo numero è approssimativamente uguale a 2^{40}, e questo significa che prima che il prodotto possa venire analizzato trascorrono almeno 40 generazioni di amplificazione. In queste 40 generazioni gli stampi più inefficienti possono facilmente essere stati migliorati.)

Esperimenti basati su diluizioni in serie, per mezzo delle quali la crescita può venire prolungata per moltissimi cicli di amplificazione, dimostrarono rapidamente la struttura dei prodotti ottimali. La loro lunghezza variava generalmente tra 150 e 250 nucleotidi. Ogni insieme di condizioni sperimentali dava un prodotto finale caratteristico e uniforme, ma c'erano tanti prodotti ottimali quanti erano gli insiemi di condizioni sperimentali. Uno dei prodotti ottimali si dimostrò essere la minivariante di Spiegelman (che nelle condizioni sperimentali di Sumper aveva mostrato di essere riproducibile). Altri prodotti ottimali erano adattati a condizioni che avrebbero distrutto la maggior parte degli RNA, come per esempio elevate concentrazioni di ribonucleasi, un enzima che degrada l'RNA in corti frammenti. Apparentemente la variante resistente alla degradazione si ripiega in maniera tale da proteggere i siti che verrebbero attaccati dalla ribonucleasi. Alcuni prodotti erano così ben adattati ad ambienti insoliti da presentare un'efficienza di replicazione 1000 volte maggiore di quella delle varianti adattate a condizioni normali.

La replicazione del DNA a doppio filamento è molto più «sofisticata» rispetto a quella dell'RNA e include dei meccanismi per riconoscere e correggere gli errori. In essa sono coinvolti venti o più enzimi. All'altezza della biforcazione replicativa una proteina separa i due filamenti progenitori; una proteina che si lega al DNA a filamento singolo mantiene separati i filamenti. Poiché la replicazione procede lungo entrambi i filamenti dello stampo in direzione 5' - 3' il processo è discontinuo per uno *(a destra nel disegno)* e continuo per l'altro *(a sinistra)*. Un promotore mobile fornisce un sito di riconoscimento per una primasi che sintetizza un blocco *(primer)* di RNA (che viene successivamente sostituito con DNA). La polimerasi III aggiunge ad uno ad uno i monomeri di DNA e allunga il filamento; la polimerasi I controlla la correttezza della sequenza, esclude i nucleotidi scorrettamente appaiati e li sostituisce con i nucleotidi corretti. Infine l'enzima ligasi salda le lacune tra i frammenti dei filamenti neoformati. In mancanza di una funzione di controllo, la replicazione del DNA non è più precisa della replicazione dell'RNA.

Queste osservazioni non lasciano alcun dubbio: i risultati di Sumper dimostravano la sintesi *ex novo*. L'uniformità dei prodotti sintetizzati *ex novo* si dimostra essere una conseguenza della selezione naturale e non di istruzioni dettagliate fornite dall'enzima di replicazione. Il dogma centrale è salvo, almeno nelle sue linee essenziali.

Ciò che è più importante, qui, è quello che questi esperimenti rivelano sui processi darwiniani. La selezione naturale e l'evoluzione, che sono conseguenze dell'autoreplicazione, operano a livello delle molecole così come a livello delle cellule o delle specie. Ciò che è veramente sorprendente, e costituisce una scoperta di grande importanza, è l'alta efficienza del processo di adattamento in un sistema di autoreplicazione così semplice.

Si potrebbe obiettare che un enzima come la Q_β replicasi, una macromolecola biologica complessa, non dovrebbe comparire in esperimenti che intendono rappresentare fedelmente la situazione prebiotica, anche se nel sistema la replicasi non rappresenta un elemento soggetto alla riproduzione o dell'evoluzione, ma semplicemente un «fattore ambientale». L'obiezione è corretta e solleva un altro importante problema.

Stampi senza enzimi

Se il funzionamento dell'RNA come stampo richiedesse sempre la partecipazione di un oggetto complesso quanto la Q_β replicasi, l'evoluzione prebiotica avrebbe richiesto procedure di ottimizzazione al di là di quelle fondate sull'autoreplicazione dell'RNA. È importante perciò stabilire quali tipi di autoreplicazione e selezione possono avere luogo in ambienti più semplici di quelli che contengono replicasi ben adattate. Possiamo

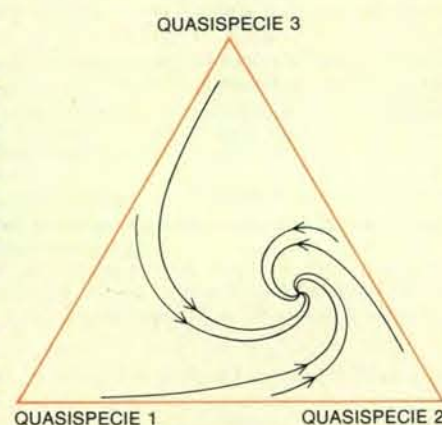
Il modello dell'iperciclo

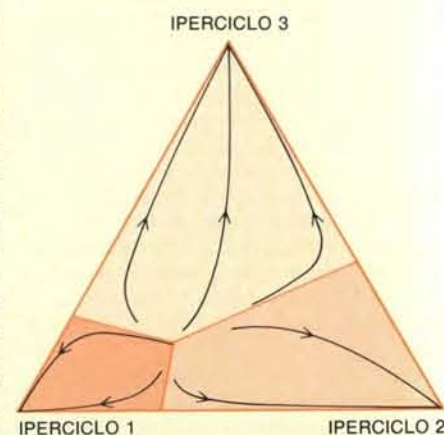
Quando la sintesi di proteine diretta dall'RNA divenne un fattore importante nell'evoluzione, si stabilì un nuovo tipo di interazione dinamica tra le molecole. Le caratteristiche principali dell'interazione possono essere chiarite da un'analisi topologica, da cui si possono trarre solo conclusioni qualitative e non quantitative.

Consideriamo un insieme di sequenze principali insieme alle loro distribuzioni di mutanti; ognuna delle sequenze principali, con la sua distribuzione associata, rappresenta (in assenza delle altre distribuzioni) una quasispecie stabile. Il contenuto informazionale totale di tutte le sequenze principali supera il limite stabilito dalla fedeltà replicativa per ogni singola quasispecie. Perché un tale insieme sia stabile e conservi la quantità totale di informazione devono essere soddisfatte tre condizioni: (1) ogni singola quasispecie deve rimanere stabile, cioè ogni sequenza principale deve competere con successo contro i propri mutanti, in modo che non si accumulino errori; (2) le diverse sequenze principali, con coefficienti selettivi diversi, devono tollerarsi a vicenda a causa di mutui accoppiamenti catalitici; (3) l'insieme accoppiato deve essere stabile, controllando la popolazione dei suoi membri e competendo come un tutto unico contro gli altri insiemi.

L'analisi topologica inizia con la definizione di uno spazio «composizionale» in cui ogni coordinata rappresenta la frazione della popolazione totale che è attribuibile a ogni quasispecie (il numero di filamenti di RNA appartenenti a una quasispecie diviso il numero totale di filamenti di RNA presenti). Un modello con tre quasispecie richiede uno spazio composizionale tridimensionale. Uno stato del sistema è caratterizzato dalle abbondanze relative delle tre popolazioni ed è rappresentato da un punto nello spazio composizionale tridimensionale. Poiché le tre abbondanze relative sono tutte positive e la loro somma è uguale a 1, ogni punto che rappresenta uno stato del sistema deve giacere su un triangolo equilatero i cui vertici si trovano sugli assi ad 1 unità di distanza dall'origine del sistema di coordinate dello spazio composizionale.

Il triangolo equilatero viene chiamato simplesso unitario. Se si considerano più di tre quasispecie, il simplesso unitario è una figura geometrica a più dimensioni. I vertici del triangolo rappresentano gli stati del sistema in cui è presente una sola quasispecie, i lati rappresentano stati in cui sono presenti due quasispecie e l'interno del triangolo stati in cui sono presenti tutte e tre le quasispecie.

Una sequenza temporale di stati è rappresentata da una «traiettoria», una curva che giace nel simplesso unitario. Sono disponibili metodi per scoprire la natura qualitativa delle traiettorie senza dover risolvere le equazioni dinamiche, che non possono essere risolte analiticamente se le quasispecie presenti sono più di due. Esistono varie classi di traiettorie possibili. In una soluzione stazionaria le traiettorie convergono in un punto stabile, e da quel momento in poi le abbondanze relative delle quasispecie rimangono stazionarie. Un'altra classe di traiettorie riflette le oscillazioni cicliche del sistema di quasispecie e una terza classe riflette un tipo di comportamento irregolare noto come «caos». Quali forme di accoppiamento permettono alle traiettorie di rimanere all'interno del simplesso unitario, implicando la coesistenza delle quasispecie?

Una traiettoria che conduce a un lato o a un vertice del simplesso unitario implica la scomparsa di una o più quasispecie. L'analisi topologica dei sistemi accoppiati in generale ha rivelato che la coesistenza, e quindi il rispetto delle tre condizioni descritte prima, richiede un tipo particolare di accoppiamento che abbiamo definito iperciclico. In un sistema iperciclico i cicli autoreplicativi sono connessi da una rete chiusa di accoppiamenti catalitici.

La classe degli ipercicli ha caratteristiche dinamiche distintive. La velocità di crescita di un iperciclo è proporzionale non alla popolazione reale di quasispecie, il che determinerebbe una crescita di tipo esponenziale, ma al numero delle popolazioni elevato a una potenza maggiore di uno. Questa crescita autocatalitica di ordine superiore al primo può venire chiamata crescita iperbolica. Gli ipercicli differiscono dai sistemi autoreplicativi darwiniani anche perché danno origine a selezioni «definitive», non più modificabili

una volta che siano state stabilite. La competizione tra ipercicli può essere analizzata topologicamente nello stesso modo della competizione tra quasispecie. Lo spazio composizionale rappresenta in questo caso l'abbondanza relativa degli ipercicli competitori, mentre le traiettorie giacciono su un simplesso unitario di ipercicli. Tutte le traiettorie conducono al vertice del simplesso unitario corrispondente alla regione in cui ha avuto inizio la competizione. In altre parole, qualsiasi competizione tra ipercicli porta alla sopravvivenza di un solo iperciclo.

La selezione «definitiva» implica che un iperciclo, una volta stabilito, non può venire soppiantato da un competitore che compaia con una abbondanza ridotta, neanche se questo è più efficiente. Ciò è una conseguenza del fatto che il vantaggio selettivo di un iperciclo dipende dalla abbondanza della sua popolazione. Questo non è vero per i sistemi darwiniani, dove un'unica molecola mutante vantaggiosa può soppiantare una popolazione precedentemente stabile. Tuttavia gli ipercicli possono evolversi ottimizzando le loro connessioni interne per mezzo di sostituzioni, inserzioni e delezioni nelle molecole informazionali. Poiché gli ipercicli stabiliscono un controllo reciproco completo tra le popolazioni accoppiate, essi crescono coerentemente.

Gli ipercicli si manifestarono in natura con la comparsa della traduzione, che introdusse un nuovo problema: la necessità di trasferire al livello genetico le informazioni sul potenziale evolutivo dei prodotti traduzionali dei geni stessi. Forse gli ipercicli esistono ancora oggi come strutture-tipo dei processi di infezione virale.

Gli ipercicli si osservano al loro livello più semplice, negli organismi attuali, quando un virus a RNA infetta una cellula. Il virus si assicura la propria replicazione preferenziale fornendo l'informazione genetica per un enzima che catalizza la moltiplicazione di quella stessa informazione. L'informazione viene fornita sotto forma di un singolo filamento «+» di RNA; questo viene tradotto dall'apparato di traduzione dell'ospite in un enzima che, con l'intervento di fattori prodotti dall'ospite, replica l'RNA e produce un filamento «—»; il filamento «—» a sua volta viene replicato in un nuovo filamento «+». Il doppio sistema a retroazione, in cui sia l'enzima codificato dallo stampo sia l'informazione costituita dalla sequenza dello stampo contribuiscono alla replicazione di quest'ultimo, è equivalente a una autocatalisi di secondo ordine.

allora considerare come possa originarsi, su basi darwiniane, la sintesi programmata delle proteine.

Questo problema deve essere affrontato sperimentalmente. Indizi importanti provengono dal lavoro recente di Leslie E. Orgel e dei suoi colleghi al Salk Institute for Biological Studies. Corti polimeri del nucleotide adenina (oligo-*A*) si formano spontaneamente quando i monomeri *A* vengono messi in presenza di lunghi polimeri del nucleotide complementare *U* (poli-*U*), che servono da stampo, anche in assenza di qualsiasi enzima o catalizzatore. La lunghezza delle catene di oligo-*A* è in media di 5 nucleotidi e può arrivare a 10. Se sono presenti ioni piombo, in funzione di catalizzatori, la resa aumenta nettamente; inoltre i monomeri sono legati in sequenza per la maggior parte (75 per cento) come nell'RNA per mezzo di un gruppo fosfato che funge da ponte tra l'atomo di carbonio 3' di uno zucchero e il carbonio 5' del successivo. Se lo stampo è poli-*C* in una miscela 50-50 di monomeri *A* e *G* attivati e in presenza di ioni piombo, i prodotti contengono *G* ed *A* in rapporto di 10 a 1; in altre parole più del 90 per cento degli appaiamenti delle basi sono corretti. Quando sono presenti ioni zinco, lo stampo poli-*C* e i monomeri *G* danno origine a catene di oligo-*G* lunghe fino a 40 nucleotidi, con una fedeltà replicativa 20 volte maggiore che in presenza degli ioni piombo. È possibile che la natura continui a «ricordare» l'origine della replicazione? Le RNA polimerasi attuali contengono tutte ioni zinco.

I dati di Orgel indicano che i polimeri ricchi in *G* e *C* devono avere offerto vantaggi speciali nei primi stadi dell'evoluzione. Solo essi possedevano una fedeltà replicativa sufficientemente alta, in assenza di replicasi ben adattate; solo essi fornivano una «adesione» sufficiente nel loro appaiamento per permettere a RNA messaggeri di una certa dimensione di venire tradotti in proteine funzionali in assenza di ribosomi, gli organelli che effettuano la traduzione nelle cellule attuali. Studi termodinamici e cinetici eseguiti da Dietmar Pörschke nel nostro laboratorio hanno posto queste conclusioni su basi quantitative. L'appaiamento *G-C* si dimostra essere 10 volte più forte dell'appaiamento *A-U*, di modo che i filamenti complementari rimangono in contatto molto più a lungo quando sono ricchi in *G* e *C*. Inoltre il legame è rafforzato cooperativamente dalle coppie vicine. Da questi dati abbiamo ricavato le regole di appaiamento per un modello evolutivo che ci permette di identificare strutture di RNA ben note (come la caratteristica struttura a trifoglio degli RNA di trasporto) come il prodotto evolutivo di processi del tipo «prova ed errore».

La conclusione fondamentale di questi studi è che l'autoreplicazione dell'RNA in sistemi *in vitro* avviene effettivamente anche senza l'intervento di enzimi sofisticati. È possibile andare avanti a considerare le conseguenze evolutive dell'autoreplicazione dell'RNA senza doversi preoccupare se questa realmente avvenne in tempi prebiotici: essa comunque avvenne.

Quasispecie di RNA

Supponete che vi sia una riserva illimitata di monomeri attivati di RNA e che la vita media delle molecole di RNA sia infinita. Che tipo di autoreplicazione avverrebbe? Qualsiasi RNA formato chimicamente senza l'intervento dell'informazione genetica verrebbe riprodotto chimicamente utilizzando le sue proprietà di stampo a una velocità proporzionale alla concentrazione istantanea di RNA. Ne risulterebbe una crescita esponenziale. Inoltre, anche se un solo stampo venisse formato inizialmente per sintesi chimica non informazionale, presto si svilupperebbe una varietà di stampi diversi a causa degli errori (mutazioni puntiformi, inserzioni e delezioni) commessi inevitabilmente nel corso della replicazione. Per questo motivo dopo ogni generazione vi sarebbe non solo una quantità maggiore di filamenti di RNA ma anche una maggiore varietà di sequenze di RNA. Cosa avverrebbe allora? Alcuni mutanti verrebbero copiati più velocemente di altri o sarebbero meno suscettibili di errori di copiatura durante la replicazione; la loro concentrazione aumenterebbe più velocemente di quella degli altri RNA. Presto o tardi questi mutanti a crescita più rapida diventerebbero prevalenti.

Si ottiene lo stesso risultato generale se si suppone che i monomeri si rendano disponibili solo lentamente, così che i polimeri in crescita debbano competere per i substrati, e se si tiene conto che i filamenti di RNA hanno vite medie finite. L'autoreplicazione è un processo competitivo; il competitore migliore è la sequenza di RNA mutante che possiede la migliore combinazione di fedeltà di replicazione, stabilità e velocità di replicazione. Questo è il filo conduttore del ragionamento che sta alla base degli esperimenti di autoreplicazione descritti precedentemente, e su cui si basa la nostra teoria dell'autoreplicazione.

La teoria della competizione nell'autoreplicazione molecolare è riassunta nella «finestra» a pagina 23. Il risultato della competizione è la «sopravvivenza» della sequenza di RNA meglio adattata alle condizioni prevalenti, che noi chiamiamo «sequenza principale», insieme a una «coda di cometa» di sequenze simili, derivate dalla sequenza principale per mutazione. Sebbene i valori dei coefficienti di velocità relativi alle condizioni chimiche primordiali non siano noti, è possibile nondimeno raggiungere delle conclusioni quantitative. Una conclusione è che esiste una condizione di soglia per l'autorepli-

cazione stabile di un messaggio genetico. Finché le circostanze non permisero il superamento di questa soglia, nessun messaggio genetico di alcun tipo poté sopravvivere.

Si può stimare la lunghezza massima dei geni disponibili nei sistemi prebiotici inserendo dei valori plausibili nell'equazione che determina la lunghezza dei geni nel nostro modello di autoreplicazione. Le lunghezze massime variano da 50 a 100 nucleotidi, una dimensione simile agli attuali RNA di trasporto. Il risultato è confortante in quanto la lunghezza è sufficiente per il ripiegamento e per la sua stabilità strutturale, ma a prima vista sembra che il messaggio genetico trasportato sia di gran lunga troppo breve per permettere la sintesi di una proteina.

Prima di mostrare in che modo questi risultati teorici siano stati confermati in laboratorio, consideriamo unicamente quali caratteristiche vengono selezionate nella competizione per l'autoreplicazione dell'RNA. Il gene selezionato è il più adatto tra quelli presenti (cioè è la sequenza principale) perché questa sequenza è quella presente in più alta concentrazione. Tuttavia la sequenza principale può costituire solo una piccola frazione del numero totale di geni. Il numero dei mutanti sarebbe estremamente elevato in condizioni prebiotiche, in quanto la cinetica chimica della maggior parte delle sequenze presenti non sarebbe molto diversa da quella della sequenza principale. Quindi il risultato della competizione per l'autoreplicazione è la sequenza principale più un enorme sciame di mutanti da essa derivati e di cui non ci si può liberare.

Definiamo questa distribuzione di mutanti «quasispecie». È la distribuzione di mutanti sotto forma di quasispecie che sopravvive nella competizione tra gli RNA autoreplicanti, e non semplicemente un'unica sequenza principale o parecchie equivalenti che rappresentano i geni più adatti nella distribuzione. L'essenza della selezione è perciò la stabilità della quasispecie. La sequenza principale diviene allora incapace di sopportare l'accumulazione degli errori; la distribuzione «va alla deriva» e infine l'informazione viene tutta perduta.

Le equazioni teoriche che descrivono la competizione autoreplicativa sono state messe alla prova e confermate con i dati degli esperimenti di clonazione con la Q_β replicasi da Charles Weissmann e collaboratori dell'Università di Zurigo. Essi hanno misurato le velocità di replicazione sul breve e sul lungo periodo e hanno studiato la competizione tra cloni mutanti e di «tipo selvatico» *(wild-type)* di RNA di Q_β. L'analisi quantitativa dei dati ha fornito coefficienti di fedeltà replicativa e vantaggio selettivo in accordo con la teoria. Gli esperimenti hanno dimostrato che anche in presenza di un apparato replicativo altamente evoluto, gli organismi si adattano a una fedeltà di replicazione non assoluta, limitando la lunghezza dei geni e sopravvivendo come distribuzioni quasi specifiche piuttosto che come geni unici.

Abbiamo citato in precedenza la crisi energetica che dovette essere superata nei primi stati della biogenesi. Discuteremo ora un ostacolo che giocò un ruolo ancora maggiore nell'evoluzione della vita: una crisi informazionale.

Errori, genotipo e fenotipo

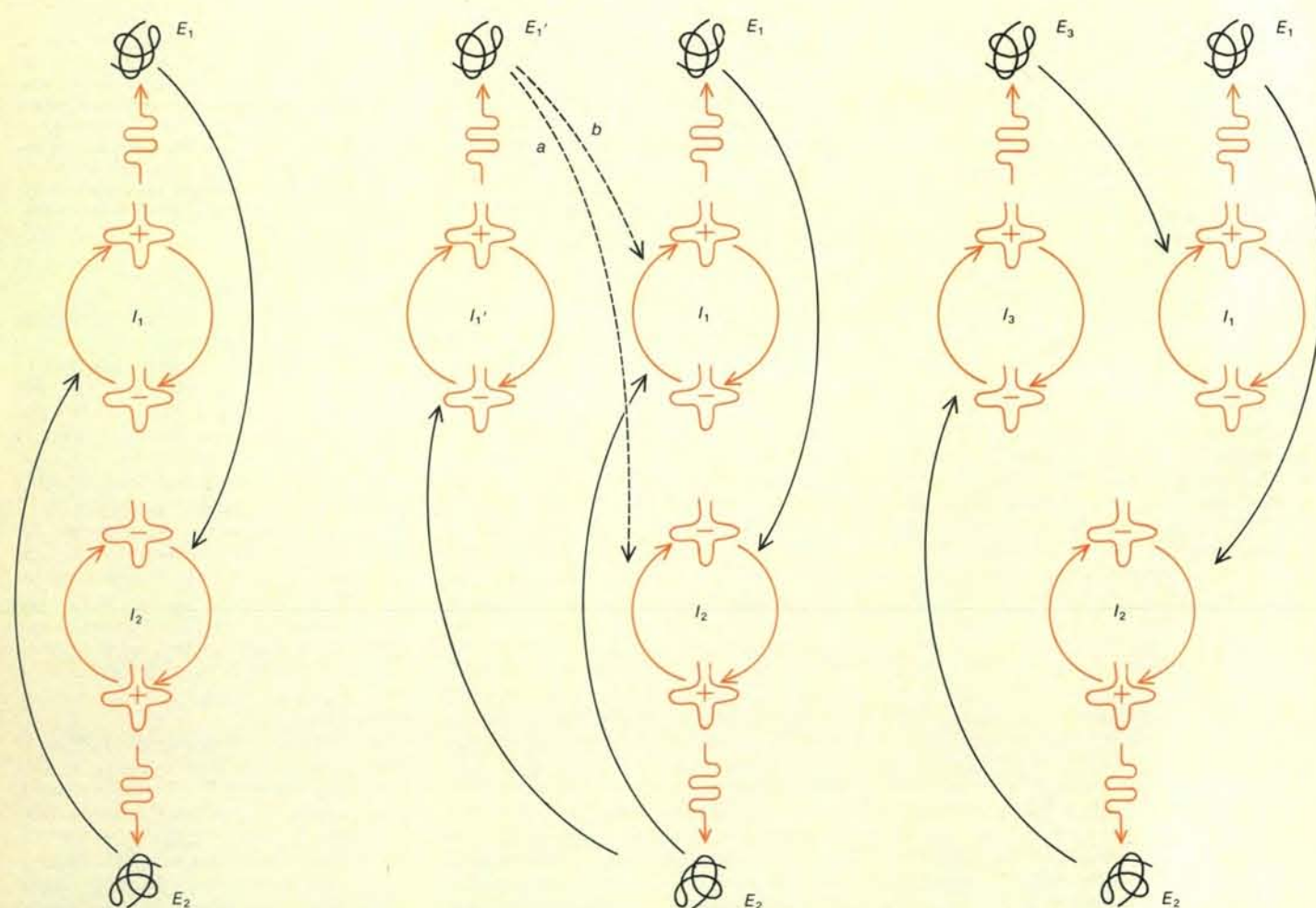
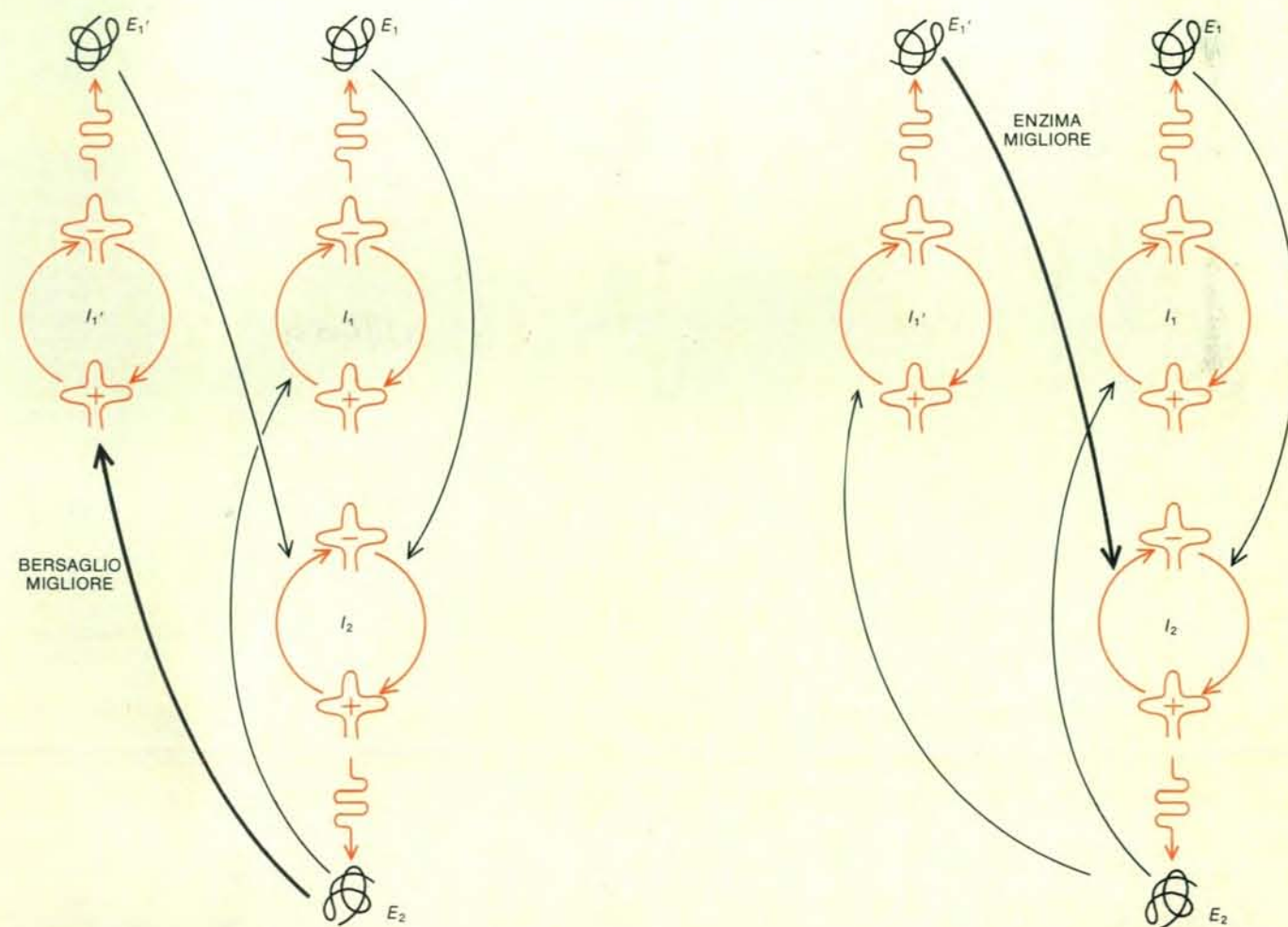
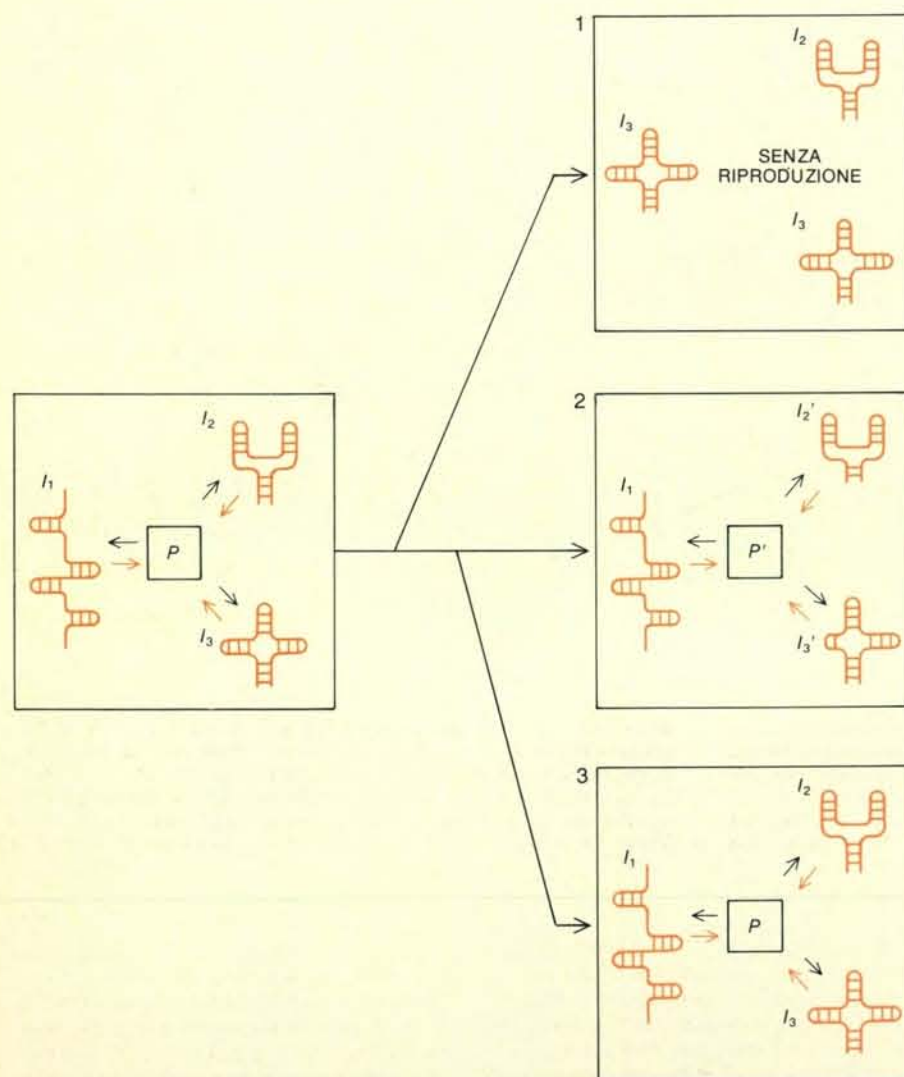
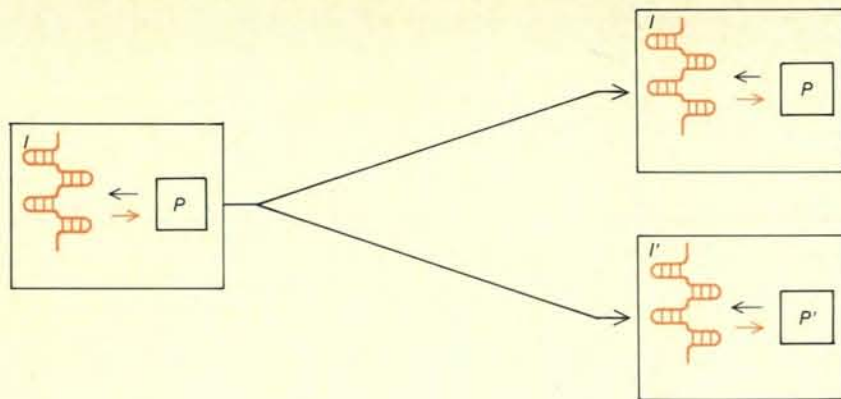
I primi sistemi molecolari darwiniani dovevano la loro capacità autoreplicativa a forze fisiche fondamentali che provocarono la formazione di coppie di basi complementari. Il vincolo dovuto alla non assoluta fedeltà replicativa limitava la lunghezza dei geni a circa 100 nucleotidi, limite raggiungibile solo dalle sequenze di RNA ricche di nucleotidi *G* e *C*. Questo limite fu alla fine superato dallo sviluppo della capacità di tradurre i geni in proteine, e quindi in sistemi enzimatici in grado di ridurre la frequenza degli errori e permettere la sopravvivenza di sequenze di RNA lunghe parecchie migliaia di nucleotidi. Questa nuova barriera è ancora riflessa nella limitata lunghezza dei geni degli attuali virus a RNA a singolo filamento, sebbene questi virus rappresentino uno stadio di sviluppo evolutivo molto posteriore.

L'ulteriore estensione della lunghezza dei geni fu possibile solo con la comparsa di meccanismi in grado di riconoscere e correggere gli errori. La distinzione tra giusto e sbagliato poté essere effettuata quando il filamento figlio appena sintetizzato rimase associato allo stampo genitore, nel qual caso «sbagliato» poté essere identificato chimicamente come «non appaiato».

Tutto ciò fu possibile quando il DNA a doppio filamento comparve sulla scena. Le DNA polimerasi sono equipaggiate con funzioni di controllo e correzione degli errori così efficaci da permettere l'estensione della lunghezza dei genomi fino a milioni di nucleotidi. Lawrence A. Loeb della University of Washington School of Medicine ha dimostrato che se una DNA polimerasi non è in grado di effettuare la correzione degli errori, ha la stessa fedeltà di replicazione di una RNA replicasi, cioè un valore che sta tra 0,999 e 0,9999.

L'invenzione del DNA rese possibile la formazione di cellule individuali in cui la divisione della cellula è sincronizzata con la replicazione del suo materiale genetico. Ora, tuttavia, doveva essere superata un'altra crisi informazionale: la replicazione ad alta fedeltà riduceva la possibilità di generare la variabilità per mezzo di mutazioni puntiformi. La nuova barriera fu superata con lo sviluppo dei processi di ricombinazione, inclusa la riproduzione sessuale, la quale ha introdotto la genetica mendeliana nel processo dell'autoriproduzione.

La prima crisi informazionale poté essere superata solo per mezzo dell'organizzazione di un meccanismo di replicazione enzimatica basato sull'esistenza di una quasispecie stabile. Questo salto evolutivo richiese la traduzione dell'informazione contenuta nell'RNA in un nuovo linguaggio di tipo funzionale: le proteine.

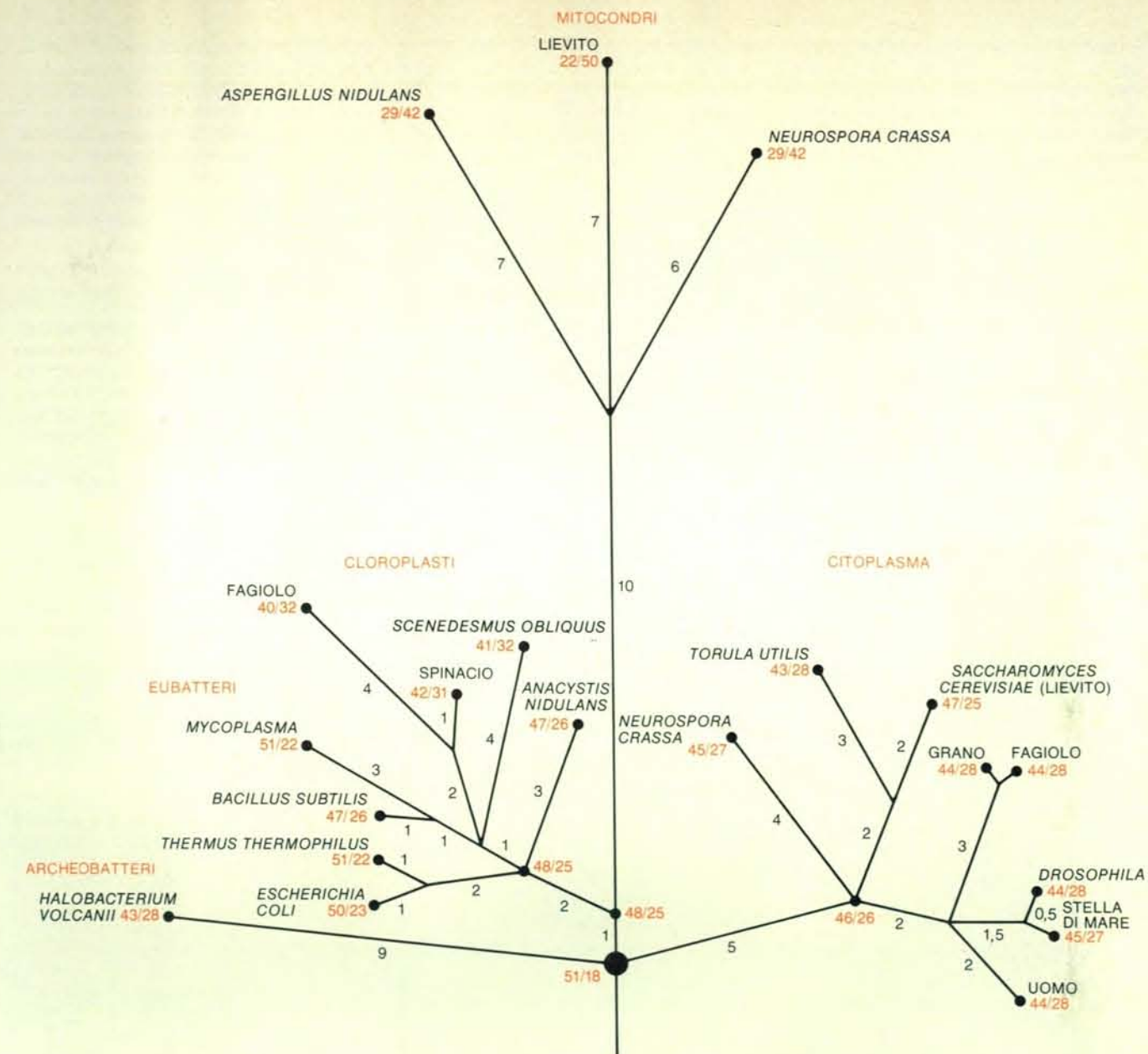
La codificazione dell'apparato di traduzione anche più primitivo richiese certamente molto più dei 100 nucleotidi circa che potevano essere conservati riproducibilmente in una sequenza principale. In qualsiasi modo si sia sviluppato il primo sistema di traduzione, deve aver ri-

L'accoppiamento iperciclico permette la cooperazione tra filamenti di RNA autoreplicanti che in caso contrario entrerebbero in competizione, provocando la scomparsa del meno adatto. Nell'iperciclo a sinistra un RNA con contenuto informazionale *(I_1)* codifica un enzima primitivo *(E_1)* che facilita la replicazione di un altro RNA *(I_2)*, che allo stesso modo facilita la replicazione di I_1 per mezzo del suo prodotto di traduzione E_2. Questo meccanismo catalitico stabilizza le concentrazioni delle molecole informazionali accoppiate a enzimi da queste codificati. Nell'iperciclo al centro emerge una sequenza mutante *(I_1')* che compete con successo con I_1; le conseguenze dipendono dal funzionamento del prodotto *(E_1')*. Se il nuovo enzima migliora l'efficienza di replicazione di I_2 *(percorso a)*, allora I_1' soppianta I_1 nel primo iperciclo. Se al contrario E_1' migliora l'efficienza di replicazione di I_1 *(percorso b)*, l'iperciclo si espande come a destra.

Le mutazioni fenotipiche e genotipiche giocarono ruoli diversi nell'evoluzione prebiotica. Una versione mutante della molecola informazionale I_1 *(I_1')* ha un effetto fenotipico *(a sinistra)* se costituisce un bersaglio migliore per l'enzima replicativo E_2 e quindi viene replicata da E_2 più efficacemente di I_1. Se E_1' catalizza la replicazione di I_2, l'iperciclo sopravvive; se ciò non avviene la mutazione è «parassita»: il mutante *(I_1')* soppianta I_1 e «spegne» l'iperciclo. Invece una versione mutata di I_1 ha un effetto genotipico *(a destra)* se non tocca l'efficienza della propria replicazione da parte di E_2, ma dà origine a un prodotto *(E_1')* la cui attività nel catalizzare la replicazione di I_2 risulta diversa dall'attività di E_1. La pressione selettiva a favore della versione migliorata di E_1 non avrebbe potuto stabilirsi in una soluzione omogenea. Infatti, prima che una tale evoluzione divenisse possibile i sistemi autoreplicativi dovettero essere isolati spazialmente in compartimenti chiusi.

La compartimentazione stabilisce una pressione selettiva sui genotipi. Supponiamo che la molecola informazionale *I* codifichi per un meccanismo *P* di traduzione e replicazione *(in alto)*. Dopo una mutazione da *I* a *I'* e la traduzione di *I'* in *P'*, il sistema può venire suddiviso nei compartimenti figli *I/P* e *I'/P' (a destra)*. Questa compartimentazione permette la selezione del gene che codifica il meccanismo più efficiente. Perché la selezione possa avvenire, la superiorità del nuovo apparato deve essere sufficiente a compensare gli errori compiuti ad ogni ciclo di replicazione. La compartimentazione di per sé, tuttavia, non è sufficiente a selezionare le informazioni migliori contenute in molecole informazionali *(in basso)*. I compartimenti figli che possiedono insiemi incompleti di componenti non sono in grado di sopravvivere *(1)*. I compartimenti che possiedono insiemi completi ma contenenti mutazioni *(2)* permettono il cambiamento evolutivo. Anche se l'insieme originale è quello ottimale *(3)*, il tasso di proliferazione richiesto per mantenere la stabilità è molto più alto di quello che sarebbe in una soluzione omogenea senza compartimenti. L'effetto netto, in altre parole, è lo stabilirsi di una condizione di limite di errore molto più restrittiva.

chiesto più informazioni di quante ne potesse fornire un primitivo sistema molecolare darwiniano. Una quantità maggiore di informazioni poteva essere stabilizzata solo con la cooperazione di diversi geni, mediata e regolata dai loro stessi prodotti di traduzione.

Se nel sistema informazionale ampliato i prodotti della traduzione sono anche i bersagli dell'evoluzione, sorge un nuovo problema. La selezione deve agire sul contenuto informazionale della sequenza nucleotidica, cioè sul suo genotipo. Il vaglio della selezione, d'altra parte, ha luogo a livello della funzione del prodotto genico: il fenotipo. Questa dicotomia genotipo-fenotipo richiede che il sistema possa trasferire informazioni verso i suoi stessi geni, un processo chiamato catalisi di secondo ordine. La specificazione «di secondo ordine» si riferisce al fatto che per riprodurre se stesso il portatore dell'informazione ha bisogno sia dell'informazione contenuta nello stampo sia dei meccanismi enzimatici codificati dallo stampo. Abbiamo chiamato i sistemi a doppio feed-back di questo tipo «ipercicli». Il termine include un'ampia classe di meccanismi autocatalitici di ordine superiore. Essi esibiscono un tipo particolare di comportamento temporale diverso da quello degli altri sistemi darwiniani.

Ipercicli: le quasi specie cooperano

L'accoppiamento iperciclico opera, oggi, quando un virus a RNA attacca una cellula. Se l'RNA virale fosse solo uno dei tanti stampi presenti nella cellula ospite, non sarebbe mai in grado di prendere il sopravvento sugli altri stampi propri dell'ospite. Esso invece specifica le informazioni per una macchina replicativa che è altamente selettiva per l'RNA virale. Molte parti del meccanismo vengono fornite dalla cellula ospite, ma la specifica relazione iperciclica assicura il successo della strategia d'attacco del virus.

Un semplice esempio aiuterà a capire il principio di funzionamento dell'iperciclo. Supponiamo che la sequenza di RNA n. 1 codifichi per un enzima 1 che contribuisce alla catalisi dell'autoreplicazione della sequenza di RNA n. 2. La sequenza 2 a sua volta codifica per un enzima 2 che contribuisce alla catalisi dell'autoreplicazione della sequenza 1. Cosa accade? La sequenza 1 ha bisogno dell'enzima 2 per autoreplicarsi e la sequenza 2 ha bisogno dell'enzima 1. Per questo motivo nessuna delle due sequenze può causare la scomparsa dell'altra competendo per la quantità disponibile di monomeri di RNA; le due sequenze sono obbligate a cooperare. A seconda delle velocità relative di numerosi gradini della catalisi, le concentrazioni prevalenti possono variare ampiamente, ma finché l'interdipendenza è conservata solo improbabili «fluttuazioni catastrofiche» o sostanziali cambiamenti delle condizioni chimiche possono rompere un iperciclo. Gli stampi e gli enzimi funzionano insieme nel processo di autoreplicazione; il prodotto proteico della traduzione serve da replicasi, da attivatore della

L'albero filogenetico dell'RNA di trasporto corrispondente al codone di inizio rivela solo piccoli cambiamenti *(numeri in nero)* nella sua sequenza nucleotidica, anche dopo un'evoluzione di miliardi di anni. La sequenza è quasi identica in tutti i mammiferi studiati finora; tra l'uomo e il moscerino della frutta *Drosophila* esistono solo piccole differenze. I numeri in colore indicano il rapporto *GC/AU*; questo rapporto è più alto vicino alle prime ramificazioni dell'albero ed è più piccolo all'estremità dei rami più lunghi. Nelle sequenze degli RNA di trasporto degli organelli cellulari chiamati mitocondri il rapporto *GC/AU* è circa 1:2, in confronto al 2:1 delle prime ramificazioni. Questo dimostra che attualmente un rapporto alto non è necessario; lo può essere stato prima poiché l'appaiamento *GC* è più stabile dell'appaiamento *AU*.

replicasi o da apparato di controllo per migliorare la velocità e la fedeltà dell'autoreplicazione.

Il comportamento cinetico di questi accoppiamenti in ipercicli è stato studiato in dettaglio ed è stato dimostrato che gli ipercicli sono le uniche strutture funzionali che permettono di superare il vincolo imposto alle quasispecie stabili dalla fedeltà di replicazione. Si dimostra che la crescita iperciclica è esplosiva, se confrontata alla crescita autocatalitica di primo ordine in un sistema con coefficienti di velocità simili. Le conseguenze della crescita iperciclica sulla selezione differiscono ancora più radicalmente.

La vita non può essersi originata con un sistema iperciclico semplice come quello descritto. I primi accoppiamenti catalitici devono essere stati deboli e complessi, con un numero molto grande di partecipanti genetici (membri delle quasispecie di RNA) e partecipanti funzionali (enzimi primitivi). Il principio ipercatalitico è nondimeno semplice: la cooperazione forzata tra geni altrimenti competitori ha permesso la sopravvivenza reciproca e ha regolato la loro crescita. Ha anche reso possibile un tipo di evoluzione più raffinata di quella accessibile alle quasispecie isolate.

In una quasispecie la competizione darwiniana valuta il successo biologico, o *fitness*, di ogni RNA mutante in base alla sua velocità e fedeltà di autoreplicazione e alla sua stabilità. Tuttavia, quando le quasispecie sono legate insieme in una struttura iperciclica, entrano in gioco nuovi criteri. Per prima cosa, la valutazione della «funzione di bersaglio» di ogni quasispecie diventa cruciale: le sequenze di maggior successo sono quelle che più rapidamente e accuratamente vengono replicate dagli appositi enzimi. Secondo, la continua introduzione di nuove sequenze mutanti permette la sperimentazione continua di nuovi accoppiamenti catalitici. La struttura dell'iperciclo evolve ogni volta che si scoprono nuovi accoppiamenti catalitici vantaggiosi.

Il principio dell'iperciclo condivide con

le quasispecie anche uno svantaggio evolutivo. Sia la competizione delle quasispecie sia la cooperazione ipérciclica valutano solo le qualità fenotipiche dell'RNA, cioè la sua stabilità e velocità di reazione. Se si potesse escogitare un modo di valutare le caratteristiche genotipiche delle sequenze di RNA, cioè il significato della sequenza stessa, la qualità degli enzimi prodotti dalla traduzione delle sequenze sarebbe soggetta a miglioramento per mezzo della selezione naturale. Secondo noi esiste solo una possibile soluzione: la separazione delle distribuzioni quasispecifiche organizzate iperciclicamente in compartimenti, che in questo modo sarebbero in grado di competere darwinianamente ed evolversi.

La transizione da una singola distribuzione quasispecifica all'organizzazione iperciclica di molte distribuzioni avvenne probabilmente per gradi e non all'improvviso. I primitivi sistemi di traduzione in una distribuzione quasispecifica diedero origine, prima o poi, a proteine casualmente presenti nel brodo primordiale. Dapprima la migliore efficienza replicativa deve avere riguardato più o meno tutte le sequenze della distribuzione quasispecifica, ma a mano a mano che le preferenze tra i prodotti di traduzione divennero più pronunciate, l'interazione reciproca tra le sequenze diventò immensamente più probabile che non il miglioramento delle singole sequenze attraverso la selezione dei prodotti di traduzione. La rete, inizialmente complessa e confusa, di interazioni divenne via via più definita quando i vantaggi derivanti dalla catalisi più specifica incominciarono a manifestarsi. E infine le differenze tra le varie interazioni stampo-catalizzatore divennero così grandi che ogni enzima si specializzò in un particolare ruolo catalitico. A questo punto la distribuzione quasispecifica originale si era differenziata in un insieme di distribuzioni quasispecifiche distinte, e il primo iperciclo poteva dirsi in funzione. Gli ipercicli nacquero altrettanto naturalmente e gradualmente quanto le quasispecie; sorsero a causa delle leggi di natura.

La compartimentazione

Oggi la vita è ovunque cellulare. Perché? Tra gli ovvi vantaggi della organizzazione cellulare vi sono la protezione dalle fluttuazioni nell'ambiente esterno e il mantenimento di gradienti di concentrazione, ma questi vantaggi non valgono a spiegare l'origine del principio di organizzazione spaziale. L'organizzazione spaziale divenne necessaria perché era il solo mezzo di risolvere l'unico problema dell'elaborazione dell'informazione che la competizione autoreplicativa e la cooperazione ipercicliche non erano in grado di affrontare: la valutazione dell'informazione contenuta nei messaggi genetici.

L'organizzazione in cellule fu sicuramente postposta finché possibile. La creazione di barriere spaziali in un sistema omogeneo deve avere introdotto problemi di difficile soluzione per la chimica prebiotica. La costruzione delle membrane, il trasporto delle sostanze attraverso di esse e la loro eventuale modificazione vengono realizzati oggi per mezzo di processi cellulari estremamente raffinati. Il raggiungimento di risultati simili nel brodo prebiotico deve avere richiesto innovazioni fondamentali.

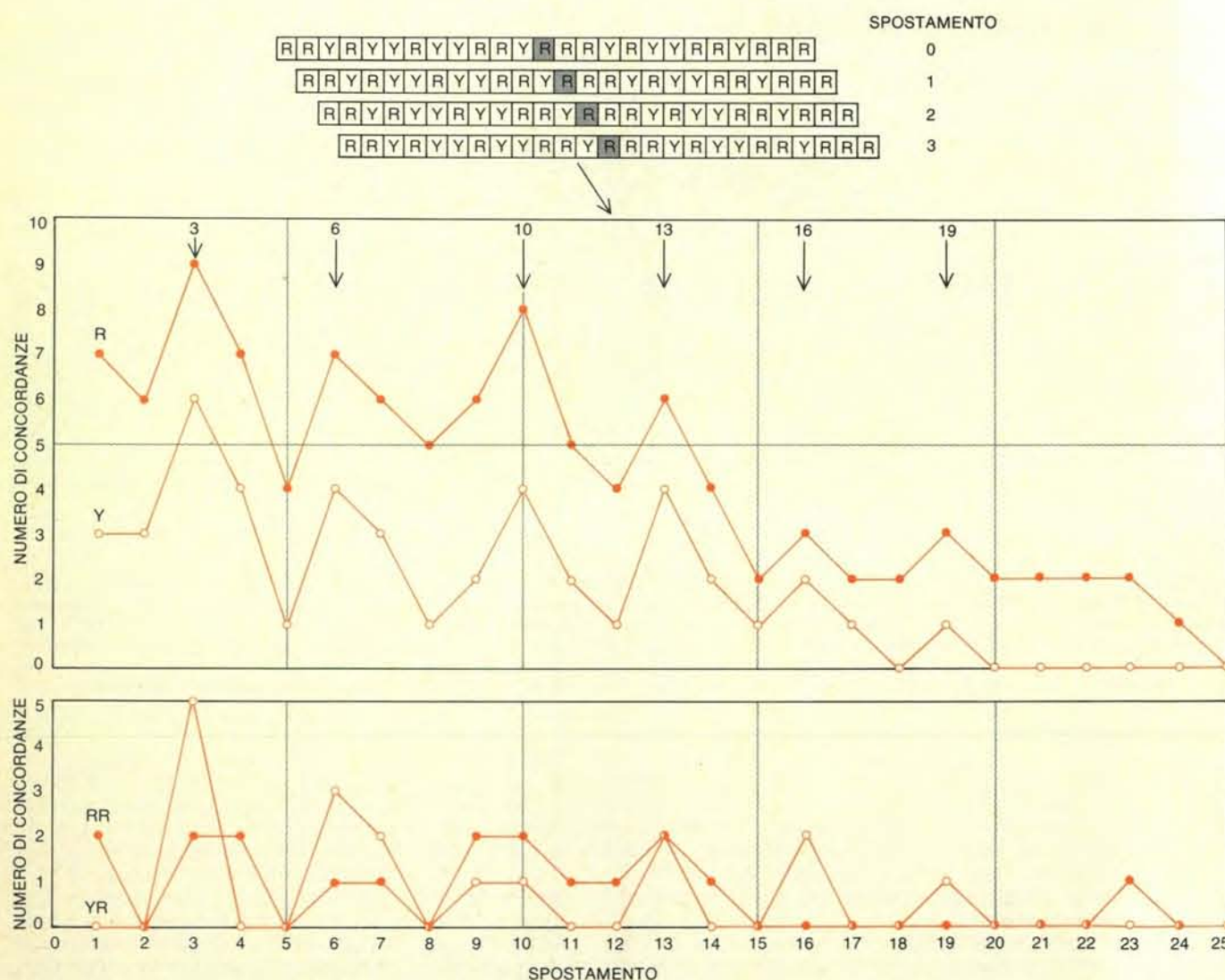
La competizione darwiniana nelle quasispecie era basata sulla selezione secondo le caratteristiche cinetiche delle sequenze; il significato del messaggio non poté essere ignorato quando entrarono in gioco l'organizzazione iperciclica degli enzimi e dell'RNA, dato che il significato determinava la forza dell'accoppiamento ciclico, tuttavia, escludeva ancora qualsiasi possibilità di valutare l'effetto del significato sul potenziale evolutivo delle sequenze, e quindi di selezionare l'informazione più utile. In un iperciclo come nelle quasispecie viene valutata solo la funzione di bersaglio dell'RNA (la sua affinità per le proteine che lo replicano) e non l'informazione genetica codificata dalla sequenza. Un iperciclo in soluzione non può scegliere i suoi prodotti di traduzione, sia che questi siano vantaggiosi, sia che siano svantaggiosi.

Siamo in grado di immaginare una sola soluzione al problema di valutare la qualità dell'informazione dei geni primitivi: la possibilità di interrompere la continuità e l'omogeneità del brodo primordiale, cioè le compartimentazione del processo evolutivo. Quando gli eventi mutazionali in un compartimento furono resi indipendenti dagli eventi contemporanei in altri compartimenti, sorse la possibilità di migliorare l'informazione genetica per mezzo dell'evoluzione darwiniana. I compartimenti più adatti poterono essere selezionati sulla base delle loro prestazioni globali, incluso il possesso di informazione genetica migliore. La possibilità di trasmettere l'informazione genetica da una generazione di compartimenti alla successiva assicurò l'evoluzione del contenuto informazionale totale.

Ora i passaggi logici dell'evoluzione prebiotica sono completi. Il raggiungimento di una qualsiasi stabilità informazionale richiese l'autoreplicazione di corte sequenze di RNA. La competizione darwiniana tra sequenze mutanti portò a una singola distribuzione quasispecifica come prodotto potenziale dell'evoluzione. In seguito, tra le sequenze mutanti si instaurò l'organizzazione iperciclica che permise a molte distribuzioni quasispecifiche di coesistere nello stesso brodo primordiale. Questo determinò un'espansione della quantità e della varietà dell'informazione presente, molto oltre i limiti imposti dalla ridotta fedeltà replicativa ai singoli geni primitivi, ma non permise la valutazione dell'informazione sulla base della funzione svolta e il suo miglioramento evolutivo attraverso la competizione. Questa opportunità fu offerta dalla compartimentazione e dalla successiva competizione tra gli svariati compartimenti.

La compartimentazione da sola non può aver soddisfatto tutte le condizioni logiche per la comparsa della vita. All'interno dei compartimenti i problemi di una limitata fedeltà replicativa e della competizione tra i geni autoreplicanti dovevano ancora essere risolti. Il solo modo finora scoperto per permettere il mantenimento di sufficiente informazione genetica per la codificazione di una quantità minima iniziale di attività enzimatica è l'organizzazione iperciclica. Gli ipercicli e i compartimenti affrontano due problemi indipendenti dell'evoluzione prebiotica. Gli ipercicli permisero la coesistenza stabile di una varietà di geni autoreplicanti e risolsero in questo modo la prima crisi informazionale. I compartimenti permisero di valutare e quindi migliorare il contenuto informazionale dei geni. In altre parole, i compartimenti risolsero la dicotomia genotipo-fenotipo.

La vita in provetta?

Se è possibile dedurre le leggi di natura che crearono la vita sulla Terra, perché non riunire i vari materiali e ricreare la vita in provetta? Chi si provasse a realizzare un simile esperimento sottovaluterebbe drammaticamente la complessità dell'evoluzione molecolare prebiotica. I ricercatori sanno solo suonare semplici melodie su uno o due strumenti della enorme orchestra che suona la sinfonia dell'evoluzione. Il ricercatore ricostruisce

L'analisi delle frequenze dei geni utilizza un metodo ideato da John Shepherd dell'Università di Basilea. Questa analisi è illustrata per una sequenza di codoni *RNY* con una base aggiuntiva (inserita alla posizione 13) e la sostituzione di un *Y* con un *R* alla posizione 25. *R* indica una purina (adenina o guanina), *N* una base qualsiasi e *Y* una pirimidina (citosina oppure uracile). Un programma al calcolatore fa scorrere la sequenza verso destra di un posto alla volta *(in alto)* e a ogni passo conta il numero di concordanze di basi singole e di coppie *RR* e *YR*. I picchi nel numero di concordanze riflettono le ripetizioni all'interno della sequenza. I picchi nelle concordanze di basi singole si riscontrano con una periodicità di 3 (con una perturbazione causata dall'inserzione), così come i picchi nelle concordanze per *YR*, come è da attendersi per un codice *RNY*. Sequenze molto più lunghe di RNA e DNA sono state esaminate per la presenza di queste correlazioni; gli effetti estranei sono stati controllati. La presenza di alcune correlazioni e l'assenza di altre sono interpretabili come un «ricordo» dell'antico codone *RNY*, confermando le argomentazioni a favore della sua antichità e i risultati ottenuti nello studio degli RNA di trasporto.

PRIMA POSIZIONE / SECONDA POSIZIONE / TERZA POSIZIONE

CODICE *N* A TRIPLETTE *GNC, GNY*

G	C	A	U
GLY	ALA	ASP	VAL

CODICE *RN* A TRIPLETTE *RNY*

	G	C	A	U
G	GLY	ALA	ASP	VAL
A	SER	THR	ASN?	ILE

CODICE *RNN*

	G	C	A	U	
G	GLY	ALA	ASP	VAL	Y
	GLY	ALA	GLU	VAL	R
A	SER	THR	ASN	ILE	Y
	ARG	THR	LYS	ILE/MET	R

CODICE *NNN*

	G	C	A	U	
G	GLY	ALA	ASP	VAL	C
	GLY	ALA	ASP	VAL	U
	GLY	ALA	GLU	VAL	G
	GLY	ALA	GLU	VAL	A
A	SER	THR	ASN	ILE	C
	SER	THR	ASN	ILE	U
	ARG	THR	LYS	MET «INIZIO»	G
	ARG	THR	LYS	ILE	A
C	ARG	PRO	HIS	LEU	C
	ARG	PRO	HIS	LEU	U
	ARG	PRO	GLN	LEU	G
	ARG	PRO	GLN	LEU	A
U	CYS	SER	TYR	PHE	C
	CYS	SER	TYR	PHE	U
	TRP	SER	«FINE»	LEU	G
	«FINE»	SER	«FINE»	LEU	A

L'evoluzione del codice genetico può essere iniziata con l'assegnazione delle quattro basi *G*, *C*, *A* e *U* nella posizione centrale del codice con sistema di lettura a triplette, ai quattro amminoacidi allora presenti in maggiore concentrazione. La tripletta era del tipo *GNC*, dove *N* può essere qualsiasi base. *U* divenne un'alternativa a *C* perché *G* può talvolta accoppiarsi con *U*, e produsse un sistema di triplette *GNY* (*Y* rappresenta *C* oppure *U*). La presenza di *U* in terza posizione rese possibile la comparsa della base complementare, *A*, in prima posizione, producendo un codice *RN* in un sistema di triplette *RNY* (*R* rappresenta *A* oppure *G*). La penetrazione di *R* in terza posizione a dare un codice *NNN* permise la penetrazione di *Y* in prima posizione nelle catene complementari, espandendo l'insieme di triplette fino al codice attuale di 64 codoni *NNN*.

il ruolo di uno strumento alla volta, nello stesso modo in cui i musicisti dilettanti si allenano a suonare il loro strumento con l'accompagnamento di un disco che riproduce l'intera orchestra, meno il loro strumento.

In questo articolo, per esempio, abbiamo descritto alcuni esperimenti *in vitro* sull'evoluzione molecolare, che hanno dimostrato come stampi di RNA con proprietà fenotipiche piuttosto sofisticate si originino in presenza della necessaria macchina enzimatica. Il passo successivo è di scoprire in che modo questa macchina si sia evoluta inizialmente. Desidereremmo ricreare tutti i suoi componenti e studiarli sperimentalmente.

Da quali amminoacidi erano composte le prime proteine sintetizzate secondo le istruzioni degli RNA? Nel codice genetico attuale ciascuna delle 64 (4^3) possibili triplette dei quattro ribonucleotidi costituisce un «codone» che informa il meccanismo di traduzione di aggiungere alla catena proteica uno dei 20 amminoacidi (o di iniziare o interrompere il processo di traduzione). Questo è sicuramente un sistema troppo complesso per aver avuto origine in un solo passaggio per cause naturali. È mai esistito un codice primitivo precursore di quello attuale? Come era strutturato? Sono queste le domande da porsi prima di poter programmare qualsiasi esperimento sull'autorganizzazione del sistema di traduzione.

Le proteine primitive dovevano essere composte da una varietà minore di amminoacidi, per cui un codice a uno o due caratteri sarebbe stato sufficiente a rappresentarli tutti. Non c'è alcun modo semplice, tuttavia, di passare da un codice a caratteri singoli, o a coppie, a un codice a triplette, perché durante una tale transizione tutti i messaggi esistenti diventerebbero privi di senso finché non fossero completamente ricodificati. Il codice genetico deve quindi essersi basato su una struttura a triplette fin dall'inizio.

La traduzione prebiotica richiedeva al codice alcune prestazioni che oggi sono svolte dal meccanismo di traduzione con accorgimenti sofisticati non correlati al codice stesso. Inizialmente dovette essere lo stesso codice a definire la direzione di lettura e la punteggiatura del messaggio (la cosiddetta *«reading-frame»*). Nel 1976 F. H. C. Crick, Sydney Brenner, Aaron Klug e George Pieczenik del Medical Research Council Laboratory di Cambridge proposero che la direzione di lettura e la punteggiatura fossero state fissate inizialmente permettendo la traduzione solo delle triplette che avevano la sequenza *RRY*, dove *R* rappresenta un nucleotide purinico (*G* oppure *A*) e *Y* un nucleotide pirimidinico (*C* oppure *U*). Notarono anche che le triplette *RNY* sarebbero servite allo stesso scopo (*N* rappresenta un nucleotide qualsiasi). La direzionalità e la punteggiatura vengono stabilite altrettanto bene da *RNY* che da *RRY*, e nel caso di *RNY* la stessa punteggiatura viene stabilita sia sul filamento che funge da stampo che sul filamento complementare.

L'evoluzione del codice

Ci sono elementi, nel codice e nel sistema di traduzione attuale, che ci permettono di stabilire che esso ebbe origine da triplette del tipo *RRY* o *RNY*? La creazione di un archivio al calcolatore di sequenze da parte di Margaret Oakley Dayhoff e dei suoi collaboratori alla National Biomedical Research Foundation ha reso possibile intraprendere ricerche su grande scala di possibili «parentele» tra polimeri biologici, e in particolare di costruire alberi filogenetici che dimostrano le correlazioni tra proteine o acidi nucleici omologhi in specie diverse. Gli RNA di trasporto si prestano particolarmente bene a questa analisi per quanto riguarda il problema delle origini. La loro funzione è di far corrispondere ogni amminoacido al suo codone. A causa di questo ruolo cruciale, la loro struttura potrebbe ancora riflettere il modo in cui si stabilì la corrispondenza tra amminoacidi e codoni.

Nella ricerca delle sequenze primitive attraverso l'analisi di quelle attuali, non è stato sufficiente elaborare una grande quantità di informazioni sulle sequenze per mezzo di un programma al calcolatore, destinato a costruire l'albero filogenetico ottimale. I criteri analitici per determinare il grado di riordinabilità in alberi di un insieme di sequenze dovettero prima essere formulati e confermati, e questo fu realizzato per mezzo di analisi topologiche eseguite grazie alla collaborazione di Andreas Dress dell'Università di Bielefeld.

Quando questi criteri e programmi del tipo di quello della Dayhoff furono applicati all'analisi di tutte le sequenze note di RNA di trasporto (circa 200), emersero due interessanti conclusioni. Primo, le sequenze di un determinato RNA di trasporto (per esempio, quello che viene utilizzato per l'inizio della traduzione) in tutte le specie studiate sembrano effettivamente essere riordinabili in una struttura ad albero che dimostra una divergenza evolutiva piccola se confrontata a quella di altri polimeri biologici. Apparentemente questo tipo particolare di informazione assai antica è rimasto relativamente inalterato durante la successiva evoluzione. Secondo, le sequenze dei diversi RNA di trasporto di uno stesso organismo riflettono una divergenza da un progenitore comune, ma non sembrano correlabili con strutture ad albero, per lo meno nei due organismi (*E.coli* e lievito) per cui possediamo abbastanza dati di sequenze per potere eseguire un'analisi statisticamente significativa.

Queste sequenze sembrano invece rappresentare una distribuzione di mutanti simili a quella di una quasispecie. In una sorta di estrapolazione all'indietro attraverso la storia delle quasispecie fino all'era di origine della traduzione, l'analisi ha identificato quelli che sembrano essere gli antenati dei moderni RNA di trasporto e ha permesso due importanti deduzioni su di essi: erano molto più ricchi in *G* e *C* che in *A* e *U*, e nella loro sequenza-prototipo (ricostruita assegnando a ogni posizione all'interno del filamento la base più comune tra le sequenze esaminate) si riesce a dimostrare una chiara reminiscenza di una struttura a triplette del tipo *RNY*.

È possibile cercare antiche informazioni genetiche anche altrove, ovunque sia possibile provare che la pressione selettiva e la deriva genetica non hanno ancora ridotto il «ricordo» delle sequenze ancestrali al di sotto del livello di rumore. John Shepherd dell'Università di Basilea ha recentemente applicato un nuovo metodo di analisi delle sequenze, per mezzo del calcolatore, adatto ai messaggi genetici lunghi. Il suo metodo misura la distanza tra ripetizioni di caratteri o gruppi di caratteri lungo la sequenza. L'informazione ancestrale può essere distinta dalle modificazioni successive. Le sue prime conclusioni, basate sullo studio di parecchi virus a DNA e di geni di batteri e organismi superiori, confermano che in questi geni moderni è ancora conservata una memoria delle antiche sequenze e che la tripletta *RNY* domina queste sequenze.

La stabilità dell'appaiamento *G-C* avvalora l'ipotesi che il codice iniziale *RNY* dovesse essere limitato ai quattro codoni *GNC*. Il significato moderno di questi codoni è *GGC* = glicina, *GCC* = alanina, *GAC* = acido aspartico e *GUC* = valina. Le simulazioni dell'ambiente chimico primordiale realizzate da Stanley L. Miller della Università della California a San Diego suggeriscono che questi amminoacidi erano tra i più comuni nel brodo primordiale. Se è una coincidenza, è certamente suggestiva.

Siamo ormai fiduciosi di poter ricostruire le sequenze degli RNA e delle proteine ancestrali. Sulla base di queste informazioni stiamo iniziando a ricostruire e risintetizzare le sequenze ancestrali, sia delle proteine sia degli RNA e a studiare i loro rapporti reciproci in un reattore a flusso continuo: in pratica, una «macchina dell'evoluzione».

Se le prime proteine erano effettivamente costituite dai quattro amminoacidi citati, esse avevano una carica elettrica negativa. In genere questi amminoacidi non avrebbero tendenza ad associarsi con molecole di RNA cariche negativamente, a meno che non esistano forze specifiche atte a stabilizzare una data interazione. Claude Hélène dell'Università di Orléans ha dimostrato che esiste una forte interazione specifica tra i gruppi carbossilato (COO^-) di amminoacidi quali l'acido aspartico e i nucleotidi *G* dell'RNA. Quindi particolari sequenze possono effettivamente stabilire strutture per contatti specifici, stabilizzabili con ioni metallici. I primi catalizzatori specifici della replicazione e della traduzione furono probabilmente queste semplici strutture, mediatrici di contatti specifici e quindi sedi di deboli attività chimiche.

Tutte queste funzioni devono essere state reclutate tra le informazioni di una quasispecie iniziale, i mutanti della quale alla fine si differenziarono quando si organizzarono in legami funzionali iperciclici. I principi guida dell'evoluzione di una tale organizzazione sono stati formulati e verificati sperimentalmente. Ora rimane solo da scoprire quali furono le strutture molecolari favorevoli.

I gusci delle nove

Le nove sono nane bianche che espellono un involucro, o guscio, quando esplodono. Il fenomeno è provocato dall'immissione nella stella di nuovo combustibile nucleare da parte di una stella compagna

di Robert E. Williams

La maggior parte delle stelle consuma no le loro scorte di combustibile nucleare a un tasso notevolmente costante per centinaia di milioni o miliardi di anni. All'incirca una volta ogni dieci anni, tuttavia, una stella della regione della Via Lattea in cui si trova il sistema solare diventa improvvisamente più luminosa di un fattore compreso fra 10 000 e un milione e, per un breve periodo, compete con le stelle più luminose del cielo: diviene una nova. La più recente nova luminosa è apparsa nella costellazione Cygnus nella tarda estate del 1975.

L'incidenza complessiva delle nove nella nostra galassia, composta da più di 100 miliardi di stelle disseminate su un disco di circa 80 000 anni luce di diametro, è difficilmente accertabile soprattutto a causa della polvere interstellare. In assenza di polvere la luminosità delle nove dovrebbe essere visibile a occhio nudo a una distanza di 25 000 anni luce. Una tipica nova in una regione fortemente oscurata dalla polvere, tuttavia, potrebbe brillare a una distanza di circa 1000 anni luce e non essere mai notata. La migliore stima dell'incidenza delle nove è stata ottenuta controllando la loro presenza nelle galassie a spirale vicine a noi. Tali indagini hanno anche dimostrato che possono essere individuati due tipi di esplosioni stellari qualitativamente differenti: nove e supernove. In una galassia tipica vi sono circa 25 nove ogni anno, ma solo circa due o tre supernove per secolo. Nessuna supernova è stata vista nella nostra galassia dal 1604. Le supernove sono decine di migliaia di volte più luminose e più ricche di energia delle nove.

I due fenomeni sono completamente differenti e non correlati. Una supernova rappresenta il breve evento finale nell'evoluzione di una stella molto massiccia. A temperature e pressioni sufficientemente elevate i nuclei degli elementi pesanti presenti nel centro della stella prendono parte a una serie di reazioni che assorbono energia dalla stella, provocando una rapida contrazione e quindi un'esplosione di rimbalzo. Nel processo, la stella viene smembrata tranne che nel nucleo centrale estremamente denso il quale rimane come stella di neutroni: il cosiddetto «resto» di supernova. Le nove, invece, non fanno parte dell'evoluzione di stelle normali: sono episodi termonucleari che hanno luogo, probabilmente periodicamente, sulla superficie di una nana bianca confinata gravitazionalmente accanto a una stella fredda molto più grande in espansione.

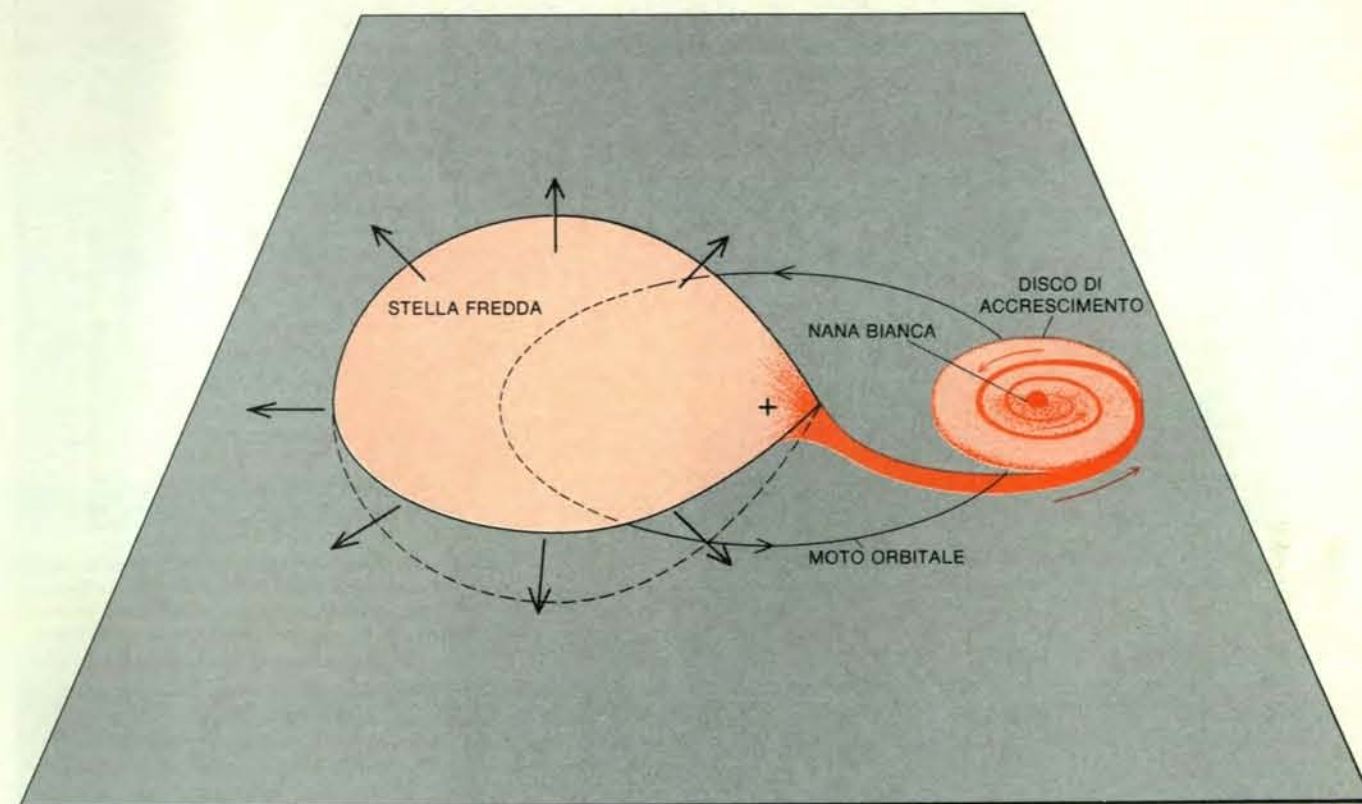
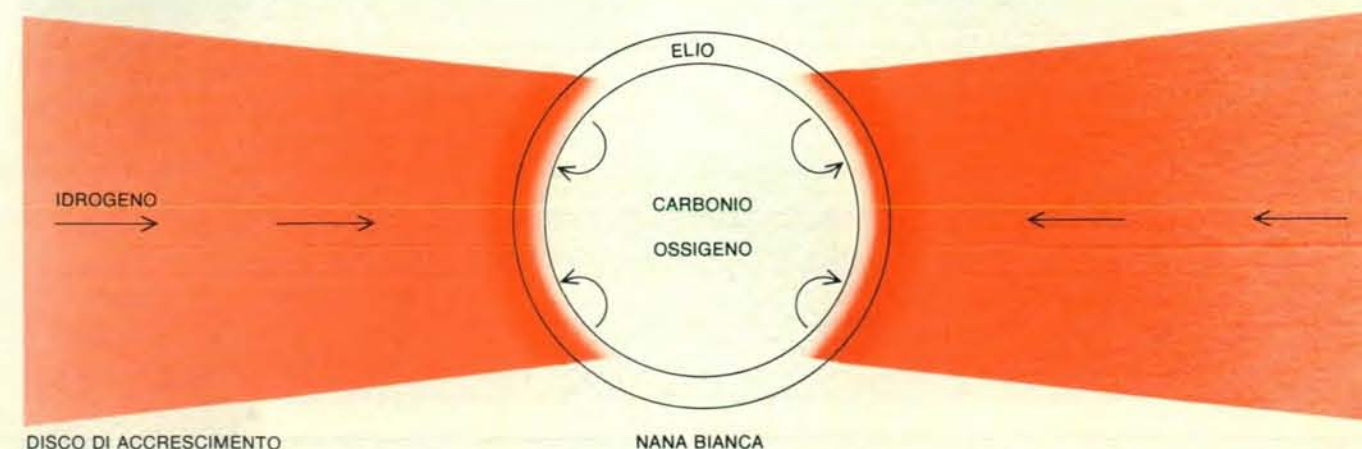
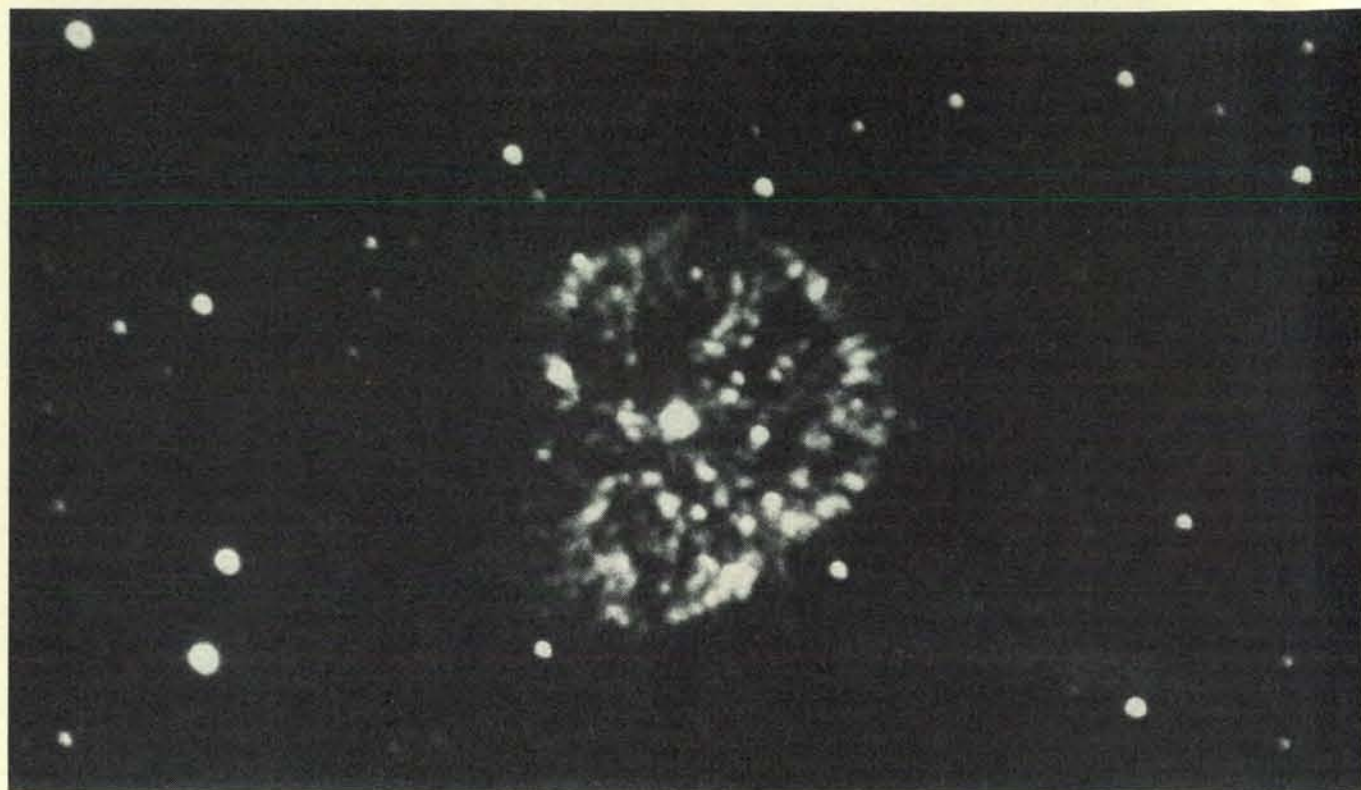
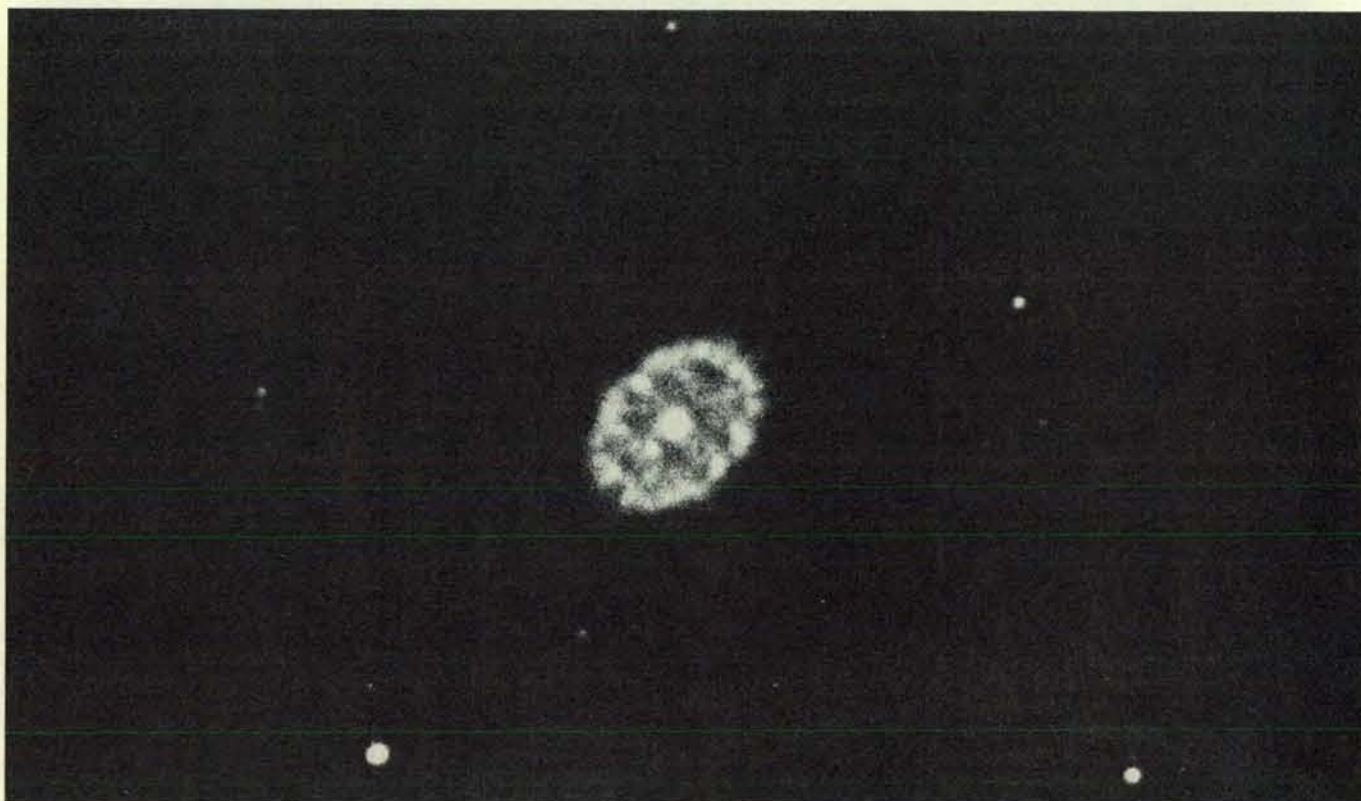
Nonostante le differenze strutturali, le nove e le supernove condividono una caratteristica: entrambe espellono materiale gassoso nello spazio. I resti delle esplosioni di supernova nella forma di nebulose filamentose possono durare per secoli. La Nebulosa del Granchio, nella costellazione del Toro, è il resto di una supernova che esplose nel 1054. Gli involucri, o i gusci, espulsi dalle nove sono più piccoli e di minor massa dei resti di supernova e possono essere visti, solitamente, solo nei pressi delle nove più luminose e più vicine.

Il tempo che impiega una nova a raggiungere la luminosità massima può variare considerevolmente, anche da due giorni a più mesi. Un grafico della luminosità apparente di una nova su un periodo di tempo è chiamato la curva di luce della nova. La curva di luce distingue due classi generali di nove, quelle veloci e quelle lente. Le nove veloci aumentano abitualmente in luminosità per un fattore superiore a 100 000 in solo pochi giorni. Il loro picco di luminosità è mantenuto per meno di una settimana e, quindi, diminuisce costantemente. Inizialmente il declino di una nova veloce è piuttosto rapido: una diminuzione di una magnitudine, o grandezza, stellare (un fattore di 2,512) ogni due giorni. Le nove lente raggiungono il loro massimo di luminosità più gradualmente e più imprevedibilmente delle nove veloci e hanno un declino molto più lento. Inoltre il loro aumento in luminosità è, approssimativamente, minore. L'energia totale liberata in un'esplosione è tuttavia la stessa per tutti e due i tipi di nove. Per quanto si può dire dalle osservazioni, le nove alla fine ritornano alla stessa luminosità che avevano prima dell'esplosione.

Inizialmente le nove vengono catalogate con il nome della costellazione in cui vengono osservate e con l'anno dell'esplosione. Così la nova del Cygnus del 1975, che per diverse notti è stata luminosa come la stella di prima magnitudine Deneb, è stata chiamata Nova Cygni 1975. Successivamente, alla nova viene dato un nome ufficiale: il genitivo latino della costellazione è preceduto da lettere o da una V (per variabile) e da un numero che indica l'ordine di scoperta della stella variabile nella costellazione. Così Nova Cygni 1975 è ora ufficialmente nota come V1500 Cygni, e un'altra ben nota nova luminosa, Nova Herculis 1934, è conosciuta come DQ Herculis.

La nostra attuale comprensione dell'esplosione di nova si può riferire alla scoperta fatta nel 1954 da Merle F. Walker, dei Mount Wilson and Palomar Observatories, che DQ Herculis è un sistema binario a eclissi, ossia una stella doppia. Walker stava studiando l'emissione di luce da DQ Herculis nello sforzo di determinare la natura dei bagliori che molte vecchie nove mostrano, quando osservò che la luminosità della stella diminuiva significativamente per circa un'ora e quindi ritornava alla sua originale intensità. Ulteriori osservazioni stabilirono che lo stesso profilo di luminosità si ripeteva ogni quattro ore e 39 minuti; la stella veniva inspiegabilmente eclissata da una compagna orbitante invisibile. La scoperta che DQ Herculis era proprio una stella doppia, col più breve periodo di rivoluzione di ogni sistema binario allora conosciuto, consentì di determinare alcune proprietà fondamentali delle due stelle.

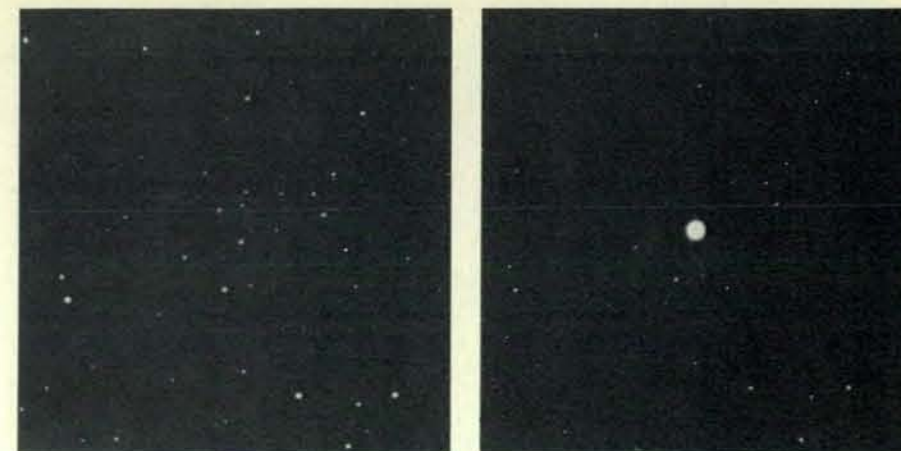
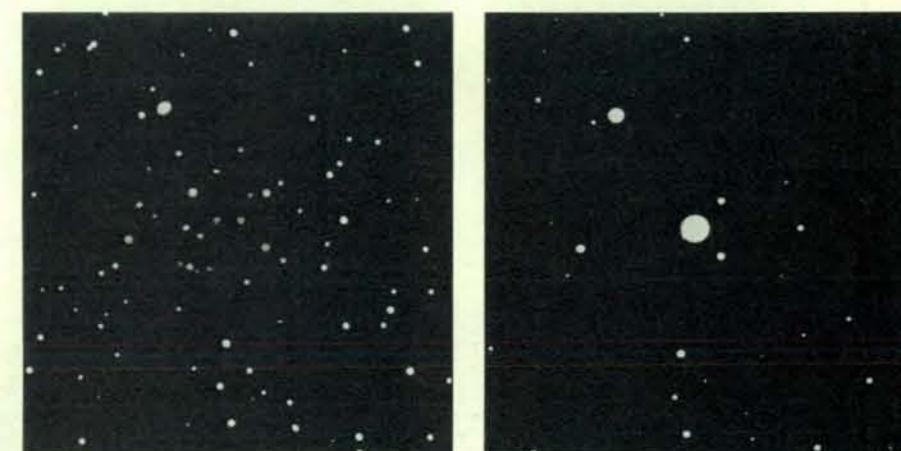
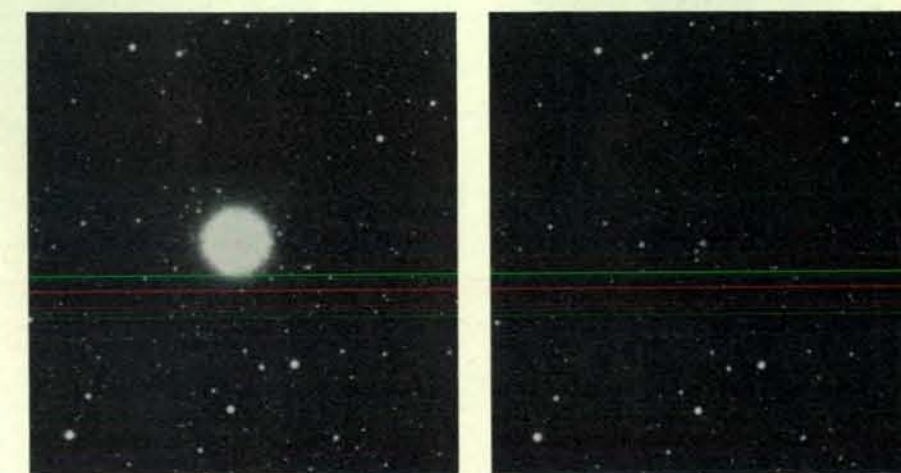
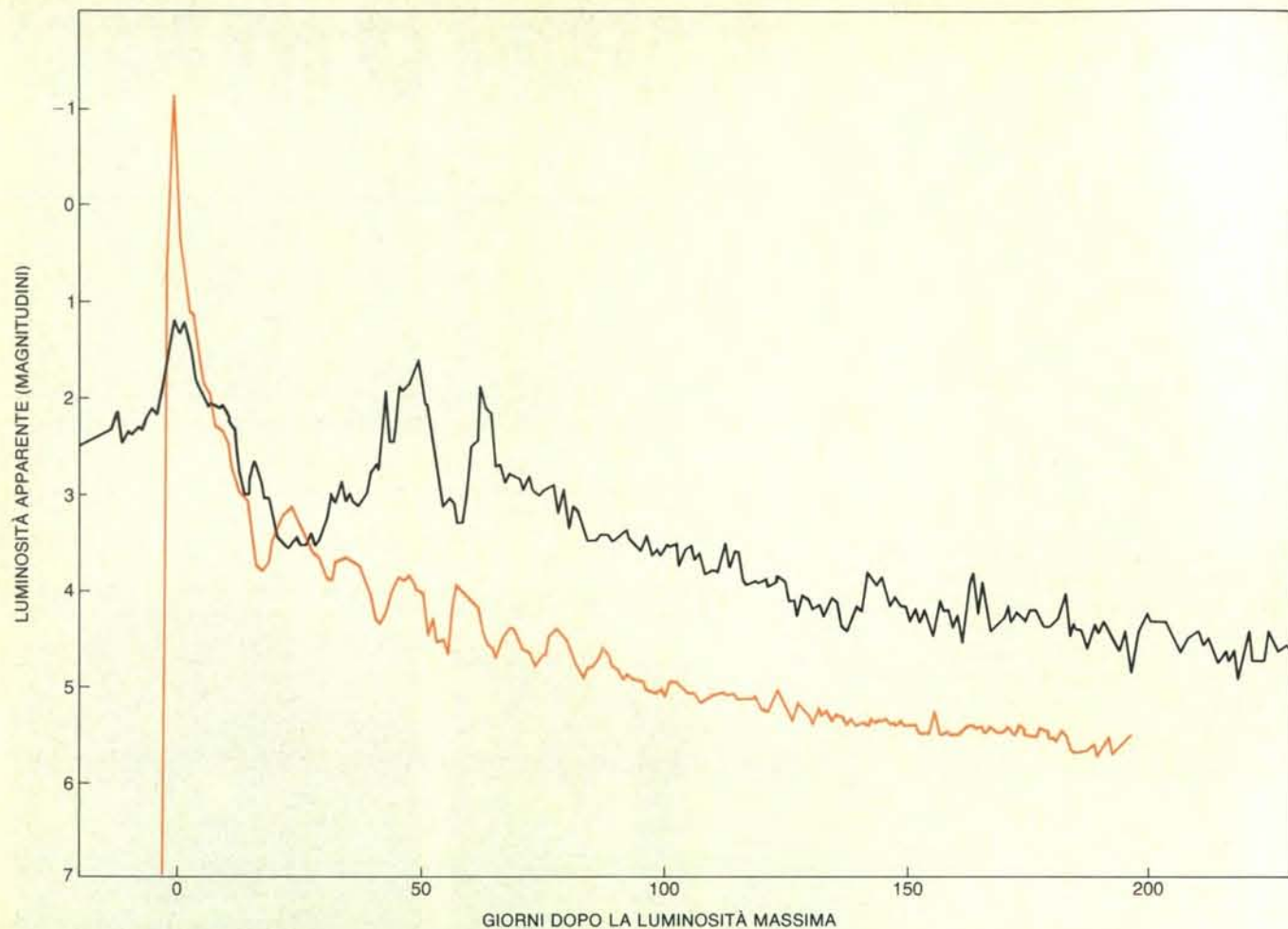
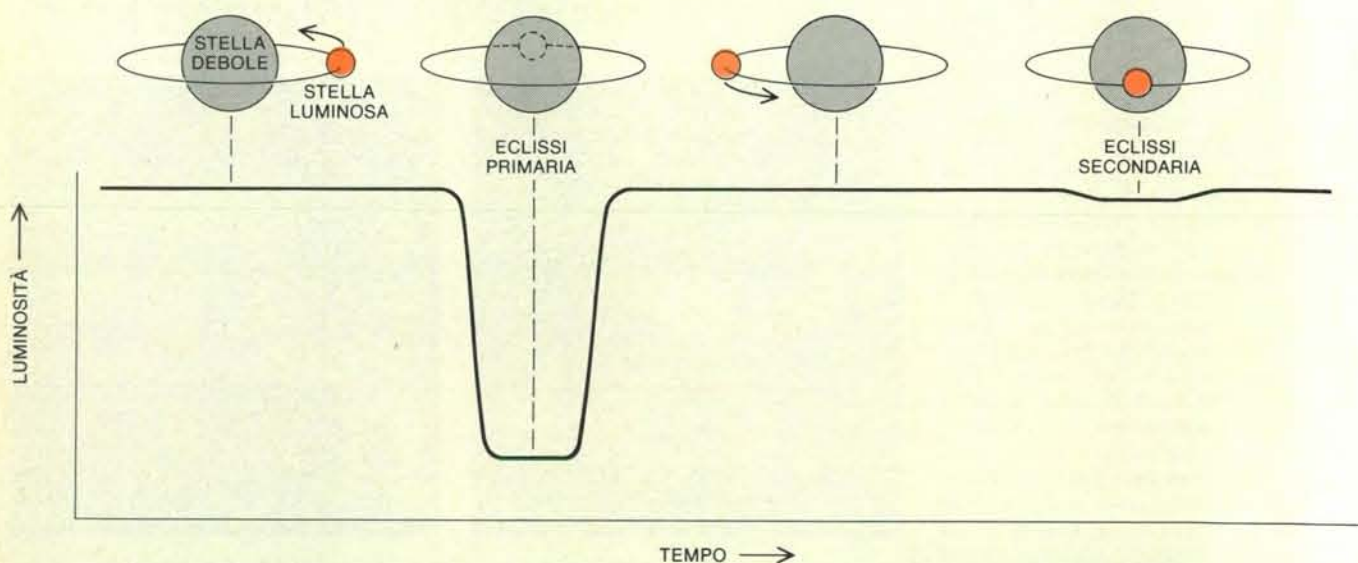
Una legge ben nota in astronomia, che può essere ricavata dalle leggi del moto e della gravitazione di Newton, stabilisce che la separazione fra due oggetti confinati gravitazionalmente in orbite l'uno attorno all'altro può essere ricavata dalle loro masse e dai loro periodi orbitali. Le masse della maggior parte delle stelle cadono in un intervallo abbastanza stretto, da circa un decimo a dieci masse solari, cosicché la distanza fra i due corpi nel sistema DQ Herculis può essere stimata con buona precisione anche senza conoscere le masse specifiche delle due stelle. A causa del brevissimo periodo del sistema era evidente che le due stelle dovevano essere molto vicine l'una all'altra, separate da una distanza di poco superiore al diametro di una stella media.

Gli astronomi trovarono questo risultato molto interessante. Era solo un caso che la nova del 1934 fosse un membro di un sistema binario molto ravvicinato? Oppure l'esplosione era un qualcosa di connesso alla grande vicinanza delle due stelle? Si iniziò una ricerca per rilevare possibili stelle compagne in altri sistemi di nove noti. Tutte le nove erano troppo lontane perché fosse possibile risolvere un sistema di stelle doppie vicine in stelle singole e così si dovette ricorrere a misu-

Nell'illustrazione è raffigurato schematicamente un sistema binario di nova visto da sopra il piano delle orbite. La crocetta indica il centro di rotazione. La nana bianca del sistema è una stella che ha convertito quasi completamente tutte le sue riserve iniziali di idrogeno ed elio in elementi più pesanti. Mancando una sorgente di energia, si è contratta fino a raggiungere un diametro pari circa a quello della Terra. La sua compagna è una stella di grandezza normale che ha esaurito la maggior parte dell'idrogeno del suo centro e ha cominciato a espandersi perché le reazioni termonucleari si sono spostate dal centro verso la superficie. Poiché la stella si espande ed è sulla strada di diventare una gigante rossa, la sua forma viene distorta dal campo di attrazione gravitazionale della nana bianca. L'idrogeno che fluisce dalla stella in espansione viene raccolto nel disco di accrescimento che si muove a vortice attorno alla stella più piccola. Alla fine la maggior parte del gas precipita lungo traiettorie a spirale e ad alta velocità sulla nana bianca, raggiungendo la temperatura della sua superficie e avviando lo stadio dell'esplosione.

L'impatto ad alta velocità del gas del disco di accrescimento, visto qui in sezione, forma uno strato caldo di confine sulla superficie della nana bianca. La materia fortemente compressa della nana bianca è in uno stato degenerato: è più simile a un solido che a un gas. Essa quindi non si espande quando viene riscaldata. L'idrogeno del disco di accrescimento si mescola con gli elementi più pesanti come carbonio e ossigeno, che possono essere trasportati dall'interno della nana bianca sulla superficie dalle correnti generate dal gas che precipita. A temperatura critica, sopra i venti milioni di kelvin, la fusione di carbonio e idrogeno avvia la reazione a catena che culmina con l'esplosione di nova.

Un guscio di gas in espansione è visibile attorno alla nova GK Persei del 1901. La nova raggiunse il suo picco rapidamente, offuscando per un breve periodo Procyon, l'ottava stella più luminosa del cielo. Il guscio, il più grande mai osservato, è molto strano, essendo non omogeneo e molto caldo. La sua temperatura di 30 000 kelvin è molto più elevata di quella che una nova può verosimilmente mantenere. Evidentemente la sua dispersione irregolare e la sua alta temperatura derivano dalla collisione del guscio con il gas del mezzo interstellare.

Il guscio attorno a DQ Herculis è stato espulso da una nova lenta che impiegò alcune settimane a raggiungere il suo massimo di intensità nel 1934, quando rivaleggiò in luminosità con Deneb, una stella della vicina costellazione del Cygnus e la diciannovesima del cielo per luminosità. La fotografia, presa sulla lunghezza d'onda emessa dall'idrogeno, è stata fatta con il telescopio da 2,3 metri dello Steward Observatory della Università dell'Arizona. A differenza del guscio attorno a GK Persei, quello attorno a DQ Herculis è simmetrico. Sebbene il gas del guscio sia ionizzato come se fosse caldo, esso è in effetti pressoché freddo: la sua temperatura è di soli 500 kelvin. L'analisi compiuta dall'autore sullo spettro del guscio mostra come questo presenti una notevole e strana abbondanza di carbonio, azoto e ossigeno. Presumibilmente questi elementi erano presenti negli strati esterni della stella nana bianca nel periodo immediatamente precedente l'esplosione.

razioni indirette. L'esistenza di sistemi binari può essere dedotta in due modi: per mezzo dell'eclissi di una stella da parte dell'altra, come nel caso di DQ Herculis, o per mezzo dell'osservazione di uno spostamento Doppler nelle righe spettrali di un tale sistema, risultante dal moto di una stella attorno alla sua compagna.

Robert P. Kraft dei Mount Wilson and Palomar Observatories studiò dieci nove vecchie e trovò che sette mostravano con grande evidenza la presenza di una stella compagna, precedentemente non rilevata. Nella maggior parte dei casi in cui poterono essere determinati i periodi orbitali dei sistemi binari, essi risultarono inferiori al giorno. Le osservazioni di Kraft non potevano rilevare un sistema binario se il piano della sua orbita era allineato quasi perpendicolarmente alla linea di vista dell'osservatore, ma sembrò ragionevole che tutte le nove studiate da Kraft non potevano rivelare un sistema rio e che le distanze che separavano le due stelle fossero generalmente comparabili con il diametro di una stella media. In ogni caso le sue osservazioni mettevano in netta evidenza che il fenomeno di una nova è collegato direttamente all'esistenza di un sistema di due stelle molto vicine.

In quei casi in cui le osservazioni di vecchie nove rivelavano un sistema a eclissi, era possibile ricavare informazioni ulteriori sulla natura delle sue stelle. Se le velocità orbitali delle stelle erano note dal periodico spostamento Doppler delle righe spettrali, le dimensioni di una stella potevano essere ricavate dalla durata di un'eclissi. Inoltre, la luminosità relativa delle due stelle può essere trovata confrontando l'intensità del sistema quando le stelle si stanno eclissando l'una l'altra con l'intensità del sistema in assenza di eclissi. Questo tipo di analisi di sistemi di vecchie nove ha portato a un altro risultato interessante e inaspettato: è comune che, in ogni sistema binario con una nova, una stella sia piuttosto calda e molto piccola. Infatti, le compagne più piccole avevano dimensioni superiori a quelle delle più piccole stelle normali. Potevano essere solo nane bianche molto dense: stelle con una massa tipica di due terzi di massa solare e grandi circa come la Terra.

Una nana bianca è una stella all'ultimo stadio del suo ciclo evolutivo. Essa ha già convertito quasi tutto il suo combustibile di idrogeno ed elio in carbonio e ossigeno per mezzo della fusione nucleare e non può più produrre energia. Mancando una sorgente di energia che la sostenga contro l'attrazione gravitazionale reciproca degli atomi che la costituiscono, essa si contrae fino a raggiungere uno stato estremamente denso e compatto, in cui la materia non può più essere compressa. La materia densissima, in una stella di questo tipo, è più simile a un solido che a un normale gas e viene descritta come degenerata. L'incompressibilità di un gas degenerato impedisce alla stella di contrarsi ulteriormente, portando a conclusione l'evoluzione stellare. Normalmente una nana bianca si raffredda lentamente per mi-

La nova DQ Herculis *(a destra)* fu fotografata nel 1934 allo Yerkes Observatory di Chicago quando la sua grandezza apparente raggiunse 1,4. Prima dell'esplosione *(al centro della fotografia a sinistra)*, la sua magnitudine era 14. La luminosità è aumentata di un fattore 100 000.

Nova Aquilae del 1918, la nova più luminosa che sia stata osservata negli ultimi cento anni, raggiunse una grandezza apparente di -1,1, diventando per alcuni giorni la seconda stella più luminosa del cielo. Nella fotografia a sinistra, fatta prima dell'esplosione, la futura nova è la debole stella di grandezza 10,6 al centro della figura. La fotografia sulla destra mostra Nova Aquilae vicino al suo picco di luminosità. Durante l'esplosione la nova aumentò in luminosità di circa 12 grandezze, o di un fattore pari a 60 000. La fotografia è dello Yerkes Observatory.

Nova Cygni del 1975, la più recente nova luminosa, fu fotografata *(a sinistra)* al Lick Observatory il 31 agosto 1975, quando essa aveva raggiunto il suo picco di grandezza a 1,8. Nella fotografia sulla destra, fatta tre mesi più tardi, la nova si è affievolita fino alla grandezza 11, diminuzione di un fattore 4800. La luminosità prima dell'esplosione non è certa perché la stella era troppo debole per essere ripresa fotograficamente. Per lo stesso motivo, tuttavia, la sua luminosità doveva essere minore di quella di una stella di grandezza 20. Di conseguenza il suo aumento di luminosità deve essere stato di un fattore di almeno 19 milioni, un record per le esplosioni di nova conosciute.

La differenza fra una nova veloce e una nova lenta può essere vista nella curva di luce di Nova Aquilae 1918 *(in colore)* e Nova Pictoris 1925 *(in nero)*. Entrambe le nove hanno gareggiato in luminosità con le stelle più brillanti. (Un cambiamento di cinque unità di grandezza corrisponde a un cambiamento di un fattore cento in luminosità.) Il rapidissimo illuminarsi di Nova Aquilae fu seguito da un rapido declino, con fluttuazioni periodiche. Nova Pictoris impiegò diversi mesi a raggiungere il suo picco di luminosità, dopo di che si affievolì più irregolarmente e più gradualmente di Nova Aquilae. Nova Pictoris fu scoperta in Sud Africa da una sentinella notturna, che la notò mentre rientrava a casa.

L'oscurarsi periodico di una stella luminosa è una prova che la stella è un membro di un sistema binario a eclissi. Con questa prova si è stabilito che i sistemi di nova sono binari e che un membro di essi è una nana bianca. Affinché sia possibile osservare l'eclissi, il piano dell'orbita del sistema deve essere visto di taglio. Si vedranno allora le due stelle eclissarsi reciprocamente a intervalli regolari. La luminosità relativa del sistema durante le due eclissi è collegata direttamente con la luminosità superficiale delle due stelle. Inoltre, il tempo affinché ciascuna eclissi divenga totale, paragonato alla durata totale dell'eclissi, può essere utilizzato per determinare la grandezza relativa delle due stelle. Nessun sistema binario è localizzato abbastanza vicino al sistema solare perché i due membri possano essere osservati separatamente.

liardi di anni e infine cade definitivamente nell'oscurità.

Sviluppando le idee di Evry Schatzman dell'Istituto di astrofisica di Parigi e di Leon Mestel della Università di Leeds, Kraft propose che l'esplosione di nova fosse causata dal trasferimento di materia dalla stella normale del sistema binario alla compagna nana bianca degenerata. A causa della sua alta densità, la nana bianca ha un campo gravitazionale molto intenso e il gas viene accelerato verso la sua superficie con velocità molto elevate. La materia che urta la superficie della nana bianca è quindi riscaldata a temperature molto alte. La temperatura può diventare così elevata da innescare reazioni nucleari sulla superficie della stella, provocando una violenta liberazione di energia.

Questa improvvisa liberazione di energia è resa possibile dal fatto che le reazioni nucleari sulla superficie della nana bianca avvengono in un gas degenerato. In un gas normale è impossibile che avvenga una liberazione improvvisa di energia perché con il crescere della temperatura il gas si espande. Tale espansione crea una caduta della temperatura del gas e il numero di reazioni nucleari, che è una funzione della temperatura, cade con essa. Questo effetto termostatico permette alle stelle normali di irradiare energia a un livello stabile per miliardi di anni.

La materia degenerata, a causa della sua incompressibilità, non si espande quando viene riscaldata. Quindi, quando avvengono reazioni nucleari, il materiale aumenta costantemente la temperatura, aumentando la velocità delle reazioni. Questo provoca un ulteriore riscaldamento del gas, cosicché il processo prosegue indisturbato, dando luogo a una serie instabile di reazioni nucleari. La produzione di energia cresce rapidamente verso livelli incredibili, culminando con una manifestazione esplosiva.

Il trasferimento di gas da una stella normale a una compagna vicina, degenerata e compatta, non è tipico delle nove. Un processo simile genera stelle a raggi X: in questo caso l'oggetto compatto del sistema binario non è una nana bianca, ma si pensa sia una stella di neutroni o un buco nero. La materia di una stella di neutroni è molto più condensata (e in un buco nero infinitamente più condensata) che in una nana bianca. Come risultato il gas che si muove a spirale nel disco di accrescimento dell'oggetto compatto accelera fino a raggiungere velocità elevatissime. A causa dell'accresciuta energia cinetica acquistata dal gas, la radiazione associata al disco è emessa soprattutto in X, invece che nel visibile.

Molti lavori teorici sulle nove compiuti nello scorso decennio hanno avuto come scopo la verifica che l'eruzione di una nova fosse un veloce processo termonucleare sulla superficie di una nana bianca di un sistema binario ravvicinato. I primi a proporre quest'ipotesi sono stati Sumner G. Starrfield dell'Arizona State University, James W. Truran, Jr., dell'Università dell'Illinois a Urbana-Champaign e Warren

Sparks del Goddard Space Flight Center. I numerosi calcoli da loro effettuati sul fenomeno delle nove sono in sostanziale accordo con molte importanti caratteristiche dell'esplosione. Il quadro che emerge dal lavoro che essi e altri hanno realizzato può essere riassunto come segue.

L'accrescimento di gas della nana bianca è continuo, guidato dalla costante espansione degli strati esterni della compagna che si evolve verso lo stadio di una gigante rossa. Il gas, che forma un disco nel piano dell'orbita della stella, è formato in gran parte da idrogeno. Quando esso ruota a spirale ad alte velocità verso la superficie della stella, si mescola con il materiale che costituisce lo strato esterno della nana bianca: elio per la maggior parte, forse con carbonio e ossigeno formatisi all'interno della stella.

Le nane bianche mancano generalmente di idrogeno perché esso si è convertito in elementi più pesanti nelle reazioni nucleari che si sono avute negli stadi di evoluzione precedenti della stella. Il costituirsi di una riserva nuova di idrogeno sulla superficie della stella è importante perché l'idrogeno, grazie alla semplicità del suo nucleo (un protone), è l'elemento che reagisce più prontamente con altri nuclei per liberare energia. È l'aggiunta di idrogeno nuovo agli elementi più pesanti già presenti nella stella che rende possibile una nuova serie di reazioni nucleari. L'inizio della reazione di fusione richiede una temperatura di circa venti milioni di kelvin. A una tale temperatura i singoli nuclei sono accelerati a velocità sufficientemente alte per superare la repulsione elettrostatica esercitata dalle loro cariche positive, permettendo ai nuclei di venire in collisione, di fondersi e liberare energia.

La temperatura necessaria si crea con l'impatto del gas, che si muove a spirale ad alte velocità, sulla superficie della stella. Il tempo richiesto perché la temperatura raggiunga il valore necessario per la fusione dipende da diversi fattori, compreso il tasso a cui la compagna che si espande rifornisce di gas il disco di accrescimento, ma è tipicamente dell'ordine delle decine di migliaia di anni. Alla fine viene raggiunta una temperatura alla quale il materiale degenerato che si trova sulla superficie della nana bianca, alimentato dall'idrogeno nuovo, comincia a sostenere la reazione di fusione.

Nelle reazioni di fusione i protoni sono catturati dai nuclei degli elementi più pesanti. I nuclei appena fusi decadono conseguentemente al tasso fissato dalla loro vita media radioattiva caratteristica. Per le reazioni postulate sulla superficie delle nane bianche il decadimento consiste abitualmente nella liberazione di un elettrone positivo (positone) e di un neutrino. La catena delle reazioni è esattamente la stessa di quella che converte idrogeno in elio al centro di una stella con una massa maggiore di quella solare. La sequenza delle collisioni, che dà luogo a una conversione netta di quattro nuclei di idrogeno in un nucleo di elio, coinvolge i nuclei di carbonio, azoto e ossigeno come catalizzatori, ed è quindi nota come ciclo CNO *(si veda l'illustrazione in basso)*.

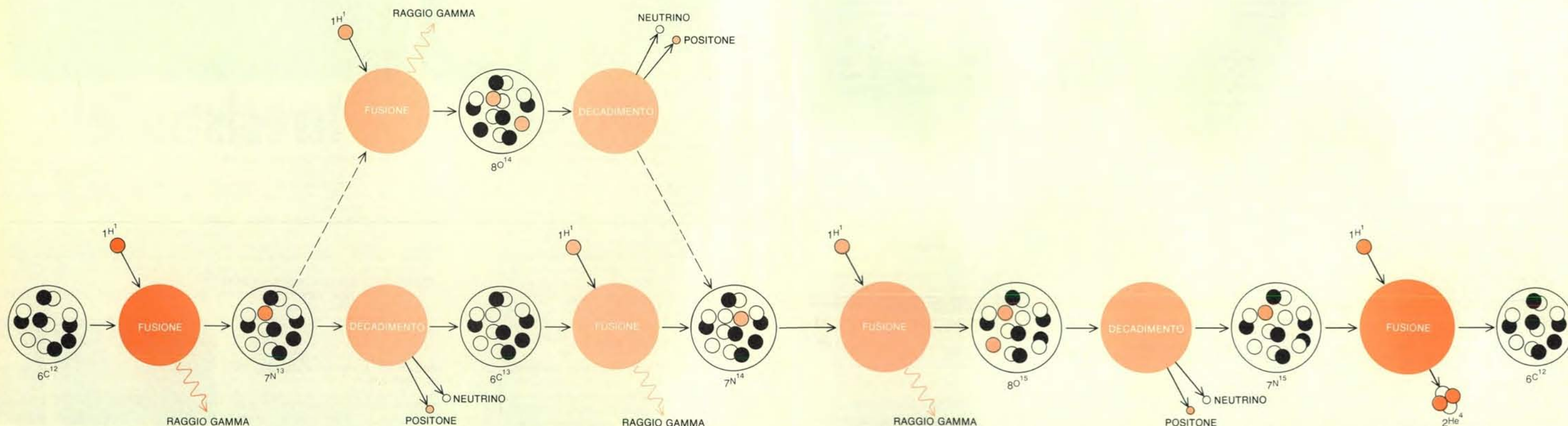
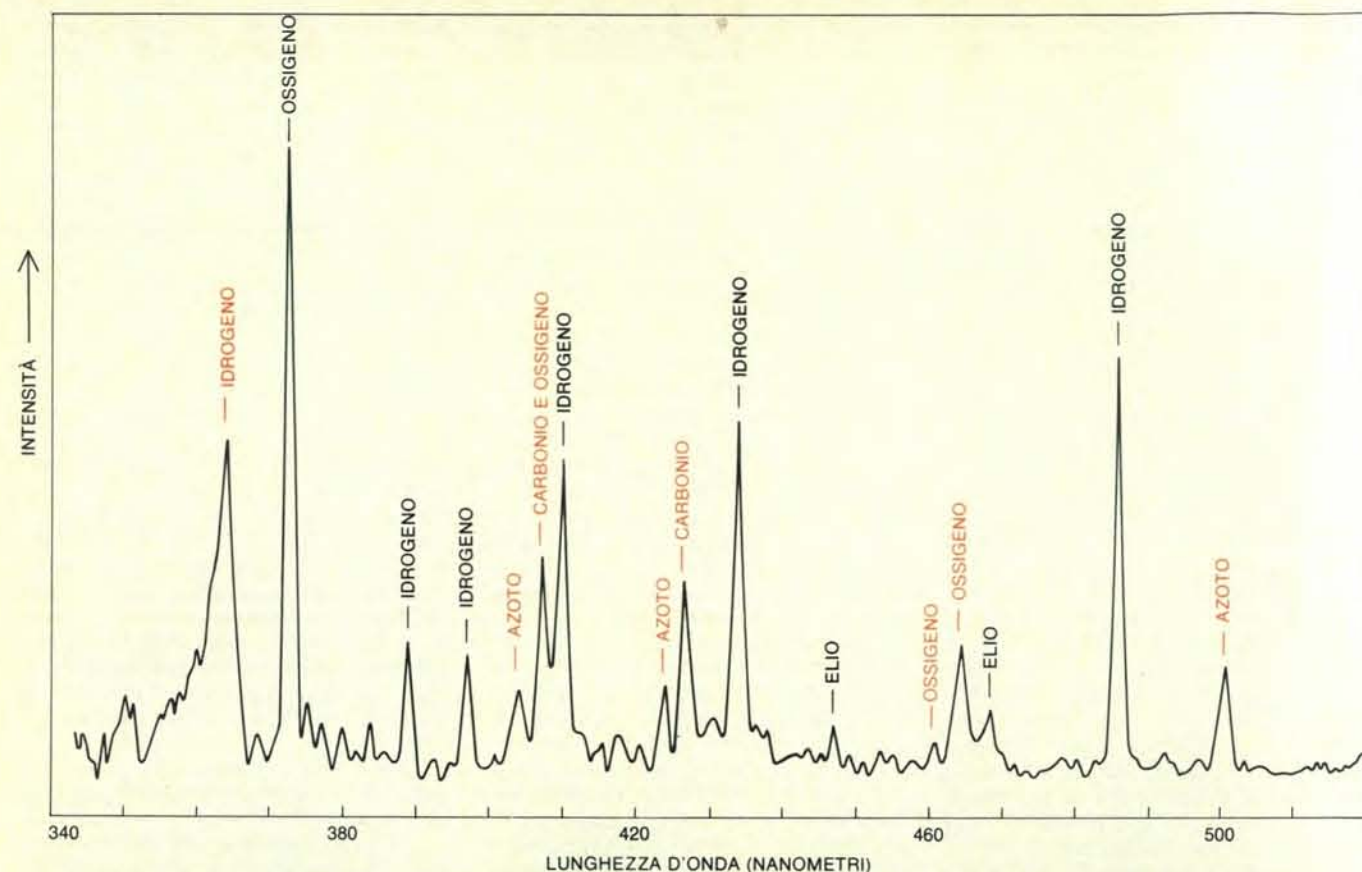
I calcoli di Starrfield, Truran e Sparks hanno dimostrato che le reazioni termonucleari del ciclo CNO si innescano effettivamente sulla superficie di una nana bianca che attira a sé, in un certo periodo di tempo, idrogeno sufficiente. I calcoli mostrano che date le condizioni che si pensano valide per sistemi binari ravvicinati, l'innescarsi delle reazioni nucleari porta effettivamente a una velocissima fuga, di origine termonucleare, della materia superficiale degenerata. Alcune caratteristiche di questi modelli teorici sono in buon accordo con il comportamento effettivamente osservato delle nove, come la forma delle curve di luce e l'energia totale generata. I calcoli mostrano anche che una delle variabili più importanti per la determinazione delle caratteristiche dell'esplosione di una nova è l'abbondanza di carbonio, azoto e ossigeno rispetto all'idrogeno. Materiale degenerato e ricco di questi elementi pesanti genera energia più velocemente e in maggior quantità di un materiale significativamente meno ricco. Infatti il modello suggerisce che una delle maggiori differenze fra nove veloci e lente sia la composizione chimica del materiale sulla superficie della nana bianca.

In base agli studi teorici, se carbonio, azoto e ossigeno sono circa 100 volte più abbondanti rispetto all'idrogeno di quanto non lo siano in una stella normale di tipo solare (ossia, se il numero di atomi di carbonio, azoto e ossigeno messi assieme rappresenta l'uno per cento del numero di atomi di idrogeno), le reazioni nucleari nel gas degenerato procedono così velocemente che la temperatura della superficie raggiunge rapidamente i 100 milioni di kelvin e in pochi minuti l'energia viene liberata in maniera esplosiva. L'esplosione sotto forma sia di radiazione sia di particelle circonda la superficie della nana bianca. Nel giro di qualche ora una frazione sostanziale dello strato di accrescimento esterno della stella viene espulso nello spazio e la stella si illumina intensamente molto in fretta, alla stregua di una nova veloce. In alternativa, il modello mostra che se gli strati degenerati della superficie hanno concentrazioni di carbonio, azoto e ossigeno non superiori a quelle di una stella normale, l'esplosione termonucleare avviene in un lasso di tempo maggiore e la stella è una nova lenta.

La previsione che le differenze nella composizione chimica possono essere largamente responsabili dei tipi differenti di esplosione di nove offre un'opportunità di vagliare la validità del modello e la rappresentazione fondamentale dell'esplosione. Forse il metodo più diretto per stabilire la composizione del gas sulla superficie della nana bianca è quello di studiare gli involucri di gas che vengono espulsi in una esplosione. Una volta espulsi, gli involucri non sono più disturbati dal sistema della nova; essi dovrebbero perciò riflettere le condizioni che hanno portato alla loro formazione.

Lo spettro delle nove durante l'esplosione indica sempre che il materiale è stato espulso rapidamente dal sistema. Gli involucri espulsi rappresentano una frazione apprezzabile del materiale che la nana bianca ha acquisito dalla compagna più grande, ammontante a circa lo 0,1 per cento della massa della nana bianca. Il materiale espulso, lanciato nello spazio alla velocità di alcune migliaia di chilometri al secondo, forma un involucro in espansione attorno al sistema binario. Tali gusci si dissolvono rapidamente man mano che espandono, scomparendo generalmente alla vista prima di essere grandi a sufficienza per essere osservati. Gli involucri espulsi da alcune nove a noi più vicine sono stati, nonostante tutto, osservati.

Allo Steward Observatory dell'Università dell'Arizona i miei colleghi e io abbiamo intrapreso un programma di ricerca per studiare i gusci delle vecchie nove. I modelli hanno suggerito che il materiale espulso dalle nove veloci dovrebbe essere ricco di carbonio, azoto e ossigeno, e i gusci attorno alle nove lente dovrebbero, al contrario, non esserlo. La composizione di un guscio e le condizioni fisiche al suo interno possono essere determinate esaminando il suo spettro. Un gas rarefatto come quello del guscio di una nova irradia a determinate lunghezze d'onda definite dalla composizione, densità e temperatura del gas. Misurando le intensità relative delle righe di emissione a diverse lunghezze d'onda, è possibile ottenere informazioni sull'abbondanza degli elementi e la temperatura dell'involucro. Il primo guscio che abbiamo analizzato spettroscopicamente è stato l'involucro simmetrico attorno a DQ Herculis, la nova lenta apparsa nel 1934. Abbiamo trovato che lo spettro del guscio, ottenuto con il telescopio da 2,3 metri dello Steward Observatory, era diverso da tutti gli altri fino allora noti. Regolarmente gli spettri delle nubi di gas che irradiano nello spazio, come quella della Grande Nebulosa di Orione, sono tutti abbastanza simili l'uno all'altro. Le nebulose brillano a causa dell'energia che ricevono dalle stelle calde a esse associate. Le stelle calde irradiano soprattutto nell'ultravioletto e, quando questa radiazione energetica illumina il gas attorno alla stella, lo ionizza, ossia strappa elettroni agli atomi. Il gas viene quindi riscaldato fino a circa 10 000 kelvin, e le collisioni fra gli ioni e gli elettroni liberi danno alla radiazione della nebulosa il caratteristico spettro di emissione.

Lo spettro del guscio di DQ Herculis era abbastanza strano: molte delle sue righe più intense non avevano un corrispettivo negli spettri di emissione di altre

La catena di reazioni nucleari che si pensa sia responsabile dell'esplosione di nova converte idrogeno in elio con lo stesso processo che ha luogo nel nucleo di una stella normale. Quattro nuclei di idrogeno, o protoni ($_1H^1$), si fondono successivamente con nuclei di carbonio, azoto e ossigeno nella sequenza di reazioni note come ciclo CNO. Le prime tre reazioni di fusione del ciclo sono accompagnate dalla liberazione di raggi gamma energetici. L'ultima reazione di fusione, nella quale un protone reagisce con un nucleo di azoto 15, è seguita dall'espulsione di un nucleo di elio, o particella alfa ($_2He^4$), e dal ricrearsi di un nucleo di carbonio 12 per ricominciare il ciclo daccapo. Alcune delle reazioni di cattura dei protoni sono seguite da reazioni di decadimento che liberano a loro volta energia sotto forma di un positone energetico (elettrone positivo) e di un neutrino. La velocità della reazione di fusione dipende sensibilmente dalla temperatura del gas. Una volta che le reazioni cominciano sulla superficie di gas degenerato della nana bianca, esse alzano la temperatura del gas, accelerando le reazioni. L'eventuale risultato può essere la reazione termonucleare inarrestabile che dà origine all'esplosione. Alcuni modelli al calcolatore suggeriscono che il fatto che l'esplosione sia lenta o veloce dipende in ogni caso dalla concentrazione di carbonio, azoto e ossigeno nel gas degenerato.

Spettro del guscio di DQ Herculis, ottenuto recentemente dall'autore e dai suoi colleghi con il telescopio da 2,3 metri dello Steward Observatory: esso è diverso da qualsiasi altro spettro di nebulosa finora conosciuta. Lo spettro del guscio di questa nova lenta consiste di una debole radiazione continua alla quale si sovrappongono righe emesse dagli elementi in differenti stati di eccitazione. La radiazione degli elementi indicati in nero è di solito predominante negli spettri delle nebulose normali, mentre le righe di emissione indicate in colore sono generalmente molto deboli o addirittura non esistono. L'intensità di emissione del carbonio, dell'azoto e dell'ossigeno mostra che l'involucro di DQ Herculis è stranamente ricco di questi elementi. La forte emissione di idrogeno a 364,5 nanometri fa ritenere che il gas sia molto freddo, circa 500 kelvin. L'emissione dell'ossigeno a 372,7 nanometri, d'altro canto, è anomala; di solito è associata a un gas alla temperatura di 10 000 kelvin.

nebulose. Il mistero dell'origine di questa strana emissione fu risolto solo quando ci rendemmo conto che una delle righe caratteristiche dello spettro, originata dall'idrogeno, richiedeva che il gas fosse molto più freddo di quello emesso dalle nebulose: circa 500 kelvin. Questo spiegava come mai molte delle comuni righe spettrali delle nebulose non apparivano nello spettro dell'involucro. Quando gli atomi emittenti furono finalmente identificati, si trovò che consistevano di ioni idrogeno, elio, carbonio, azoto e ossigeno. Questo fatto era nuovamente sorprendente perché sono necessarie alte energie per la ionizzazione degli atomi e alte energie sono abitualmente associate ad alte temperature.

La paradossale ionizzazione di un gas freddo può essere spiegata se il guscio della nova non è in equilibrio, ma subisce ancora gli effetti dell'esplosione. È noto, dalle osservazioni effettuate subito dopo l'esplosione di una nova, che il guscio emesso è fortemente ionizzato e caldo, con temperature che superano i 15 000 kelvin. Col passare del tempo la superficie della nana bianca si raffredda, non emettendo più l'intensa radiazione ultravioletta che in origine riscaldava e ionizzava il guscio. A volte, nel guscio in espansione, sono disponibili sorgenti alternative di energia. Se l'involucro si sta espandendo in una regione in cui il debole gas dello spazio interstellare è relativamente denso, le collisioni fra esso e il gas interstellare possono essere altamente energetiche. Tali collisioni possono spiegare lo strano aspetto del guscio di GK Persei, una nova brillante del 1901, che lasciò uno dei più grandi e caldi involucri conosciuti. Il guscio di DQ Herculis, d'altro canto, si sta evidentemente espandendo senza interferenze con il gas interstellare; esso giace al di fuori del piano centrale della nostra galassia, dove si trova la maggior parte del gas interstellare. Libero di espandersi senza alcuna addizionale immissione di energia, l'involucro di DQ Herculis si è raffreddato regolarmente. Gli elettroni che erano stati tolti agli atomi vengono ora ricatturati dagli ioni. Il tempo richiesto affinché gli elettroni siano completamente ricatturati, però, è sostanzialmente più lungo di quello che impiega la temperatura del gas a diminuire. Il risultato è che il guscio si è ora raffreddato molto al di sotto della temperatura originaria e la maggior parte degli elettroni non sono stati ancora ricatturati; di qui l'apparente paradosso di un gas freddo che è ancora fortemente ionizzato.

NOVA	ANNO DELLA ESPLOSIONE	GRANDEZZA MASSIMA	CLASSE DI VELOCITÀ	AMPIEZZA ANGOLARE DEL GUSCIO (SECONDI D'ARCO)	TEMPERATURA DEL GUSCIO (KELVIN)	ELEMENTI ABBONDANTI NEL GUSCIO
T AURIGAE	1891	4,2	LENTA	20	≲3000	ELIO, AZOTO E OSSIGENO
GK PERSEI	1901	0,2	VELOCE	75	>25 000	AZOTO
V476 CYGNI	1920	2,0	VELOCE	10		
RR PICTORIS	1925	1,2	LENTA	25	15 000	ELIO E AZOTO, PROBABILMENTE NEO
DQ HERCULIS	1934	1,4	LENTA	20	500	CARBONIO, AZOTO E OSSIGENO
CP PUPPIS	1942	0,2	VELOCE	15	≲1000	AZOTO
V533 HERCULIS	1963	3,0	VELOCE	5		
T PYXIDIS	1890 1944 1902 1966 1920	6,6	LENTA	10	~10 000	NESSUNO

L'autore e i suoi colleghi stanno studiando otto nove con gusci osservabili. Sebbene vi siano prove che tutte le nove espellono gas durante l'esplosione, relativamente poche producono involucri rilevabili. La maggior parte dei materiali gassosi espulsi è troppo debole per esser vista in fotografia perché i gusci tendono a scomparire prima di raggiungere una ampiezza che li renda osservabili. I gusci si formano attorno sia alle nove veloci che a quelle lente e mostrano temperature molto diverse fra loro. Con l'eccezione della ricorrente T Pyxidis, le nove sembrano formare gusci con una caratteristica in comune: essi sono molto ricchi, rispetto a una stella normale, di elementi pesanti.

La nova ricorrente T Pyxidis ha subito cinque esplosioni dal 1890, la più recente delle quali nel 1966. Le nove ricorrenti aumentano la luminosità durante l'esplosione solo di un fattore 1000, una crescita considerevolmente minore di quella di una nova normale. È stato scoperto solo di recente che anche le nove ricorrenti espellono gusci osservabili (come si vede in questa fotografia stampata in negativo di T Pyxidis, ottenuta elettronicamente con iconoscopio del telescopio da quattro metri del Kitt Peak National Observatory). Il guscio è stato scoperto nel 1978 da Harvey R. Butcher del Kitt Peak, D. A. Kopriva e dall'autore mentre studiavano gli involucri delle nove.

Quando abbiamo portato a termine l'analisi dello spettro del guscio, abbiamo trovato che la quantità di carbonio, azoto e ossigeno messi assieme è superiore di un fattore cento a quella delle stelle comuni. Secondo i modelli convenzionali, a tali quantità dovrebbe corrispondere una nova veloce. Poiché DQ Herculis era una nova lenta, tuttavia, le analisi spettrali e i modelli sono in evidente conflitto.

Si deve fare molta attenzione nel generalizzare ad altre nove i risultati relativi a un singolo oggetto e, perciò, John S. Gallagher della Università dell'Illinois a Urbana-Champaign e io stiamo ora studiando altri sistemi di nove con gusci estesi. Abbiamo finito il lavoro per quel che riguarda altri due oggetti, gli involucri che circondano la nova lenta RR Pictoris, che esplose nel 1925, e T Aurigae, del 1891. Gli spettri mostrano che il guscio di RR Pictoris è sostanzialmente arricchito di elio e azoto; quello di T Aurigae è arricchito di elio, idrogeno e ossigeno. Gary J. Ferland e Gregory A. Shields dell'Università del Texas ad Austin hanno condotto uno studio analogo di V 1500 Cygni, la nova veloce del 1975, e hanno trovato che il suo guscio, ancora troppo piccolo per essere risolto fotograficamente, è molto più ricco delle stelle normali di elementi pesanti inclusi carbonio, azoto e ossigeno.

Numerosi altri gusci di nove, alcuni dei quali nel cielo dell'emisfero australe, sono accessibili all'analisi e se ne sta studiando la composizione. Sulla base dei risultati disponibili fino a ora sembra che la maggior parte delle nove, lente e veloci, siano più ricche del normale di svariati elementi pesanti. L'origine di questo fatto non è nota, ma è possibile abbia a che fare con l'interno delle nane bianche, che sono molto ricche di carbonio e ossigeno.

La presenza di grandi quantità di elementi pesanti nelle nove veloci non è per niente sorprendente. La scoperta che i gusci attorno alle nove lente DQ Herculis, RR Pictoris e T Aurigae sono molto ricchi di elementi pesanti, tuttavia, non si accorda con i calcoli teorici. In particolare l'involucro di DQ Herculis contiene quantità di carbonio, azoto e ossigeno che, se fossero state presenti prima dell'esplosione, avrebbero dovuto generare energia sufficiente per fare di DQ Herculis una nova veloce.

La difficoltà di rinconciliare l'abbondanza di certi elementi nei gusci delle nove lente con le previsioni dei modelli del ciclo CNO ha portato alcuni teorici a supporre che non sempre l'esplosione di una nova debba coinvolgere il ciclo di reazioni CNO. L'abbondanza di certi elementi può essere spiegata se l'esplosione di una nova è stata generata da una reazione nucleare capace di sintetizzare carbonio, azoto e ossigeno e altri elementi direttamente dall'idrogeno e dall'elio, cosa che il ciclo CNO non fa. Vi sono, per esempio, reazioni termonucleari in cui tre atomi di elio (particelle alfa) si fondono per produrre carbonio 12, dal quale si possono formare altri elementi più pesanti con un'ulteriore cattura di protoni. Questo processo «alfa-triplo» è conosciuto come un'importante sorgente di energia nelle giganti rosse. Probabilmente c'è elio a sufficienza sulla superficie degenerata di una nana bianca per permettere il verificarsi di questa reazione e sembra così possibile che alcune esplosioni di nova siano dovute al processo alfa-triplo piuttosto che al ciclo CNO basato sull'idrogeno.

Una prova ulteriore della fusione dell'elio in alcune esplosioni di nova può essere dedotta confrontando l'abbondanza di elio nei dischi di accrescimento prima dell'esplosione, con quella nei gusci delle nove dopo l'esplosione. Recenti studi di vecchie nove compiuti allo Steward Observatory hanno dimostrato che l'ammontare di elio nei gusci espulsi è inferiore a quello comunemente trovato nel gas dei dischi di accrescimento attorno alle nane bianche. Una logica spiegazione per questa diminuzione del contenuto di elio del gas è che le reazioni del processo alfa-triplo che sono avvenute durante l'esplosione hanno convertito l'elio in carbonio e altri elementi.

Nonostante incertezze in alcuni dettagli, oggi vi è un accordo generale sul fatto che le esplosioni di nova sono causate da reazioni termonucleari sulla superficie delle nane bianche nei sistemi binari ravvicinati. Il processo è iniziato col trasferimento di materiale da una compagna in espansione. La materia precipita sulla superficie degenerata di una nana bianca a velocità talmente elevate che la superficie si riscalda alla temperatura di venti milioni di kelvin, necessaria per innescare reazioni nucleari esplosive. I modelli di nova basati sulle reazioni del ciclo CNO hanno permesso di spiegare con successo molte delle caratteristiche che sono state osservate in una nova. Recentemente, però, nuovi dati sulla composizione di vecchi gusci stanno portandoci a modificare alcune delle vecchie idee sulla natura dell'esplosione.

Una teoria unificata delle particelle e delle forze

In un raggio di 10^{-29} centimetri il mondo può essere un luogo semplice, con un solo tipo di particelle elementari e una forza importante. Se la teoria unificata proposta è corretta, tutta la materia è instabile

di Howard Georgi

Non ci può essere niente di più semplice di una particella elementare: è un frammento invisibile di materia, senza struttura interna e senza forma o dimensioni rivelabili. Ci si potrebbe aspettare altrettanta semplicità nelle teorie che descrivono tali particelle e le forze con le quali esse interagiscono; perlomeno, ci si potrebbe aspettare di poter spiegare la struttura del mondo con un numero minimo di particelle e di forze. Giudicandola con questo criterio di parsimonia, a una descrizione della natura che si è evoluta negli ultimi anni si può attribuire un successo ragionevole. La materia è costruita di soli due tipi di particelle elementari: i leptoni, quale l'elettrone, e i quark, costituenti del protone, del neutrone e di molte particelle correlate. Tra le particelle elementari agiscono quattro forze fondamentali. La gravitazione e l'elettromagnetismo sono da tempo familiari nel mondo macroscopico; la forza debole e la forza forte si osservano solo in eventi subnucleari. In linea di principio questo insieme di particelle e di forze potrebbe spiegare l'intera gerarchia osservata di strutture della materia, dai nuclei degli atomi alle stelle e alle galassie.

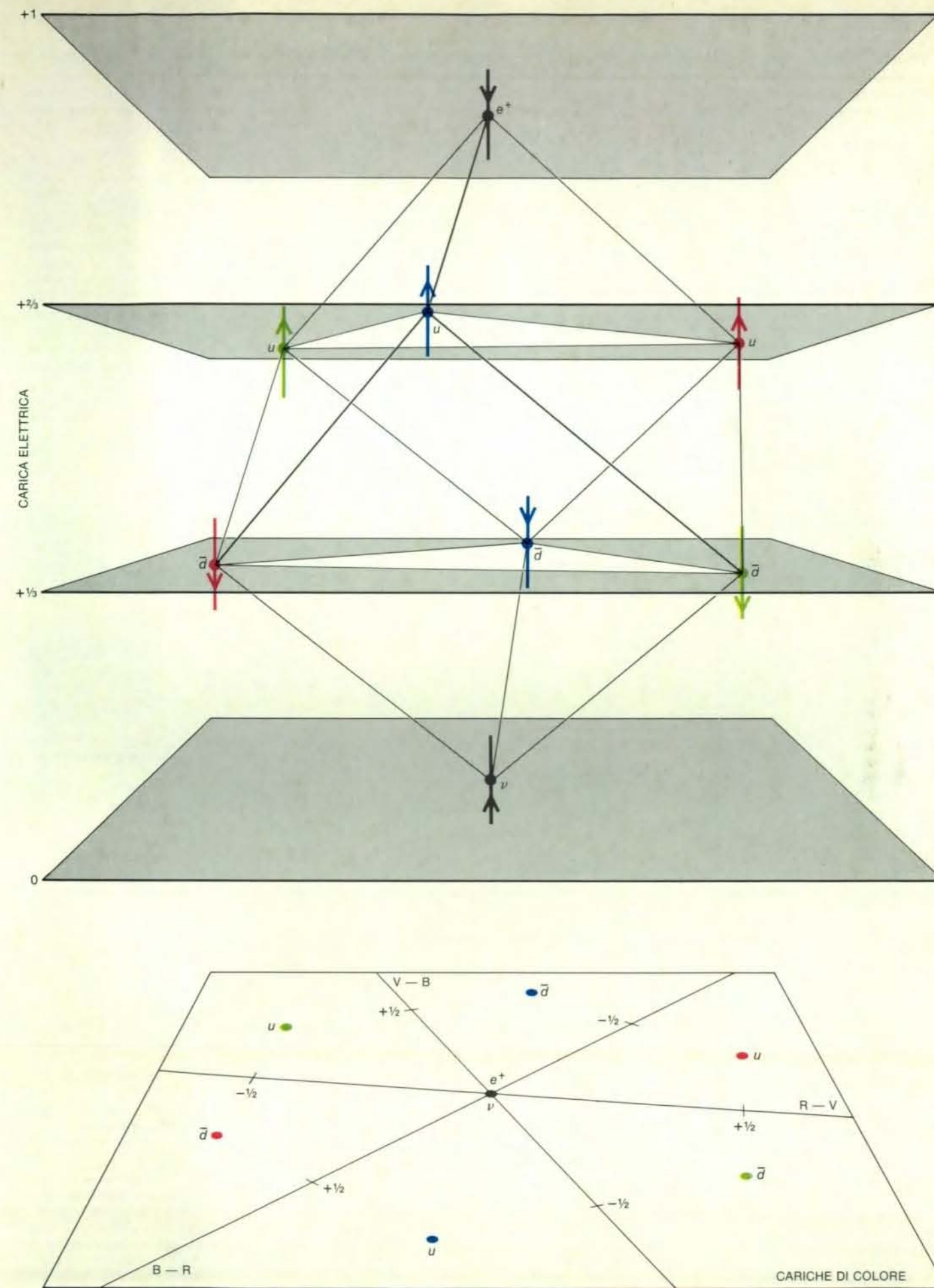
Una conoscenza della natura a questo livello di particolari è un risultato notevole; ciononostante è possibile immaginare come potrebbe essere una teoria ancora più semplice. L'esistenza di due diverse classi di particelle elementari non soddisfa pienamente; idealmente dovrebbe essere sufficiente una sola classe. Allo stesso modo, pare una inutile complicazione l'esistenza di quattro forze; una sola forza potrebbe spiegare tutte le interazioni delle particelle elementari. Una nuova, ambiziosa teoria, promette oggi almeno una parziale unificazione lungo queste linee. La teoria non comprende la gravitazione, che è di gran lunga la più debole delle forze e può essere fondamentalmente diversa dalle altre. Tuttavia, se si esclude la gravitazione, la teoria unifica tutte le particelle e tutte le forze.

Il primo passo nella costruzione della teoria unificata fu la dimostrazione che le forze debole, forte ed elettromagnetica si potrebbero descrivere tutte con teorie dello stesso tipo generale. Le tre forze rimasero distinte, ma si potè vedere che funzionavano con lo stesso meccanismo. Nel corso di questo sviluppo venne scoperta una profonda connessione tra la forza debole e l'elettromagnetismo, una connessione che suggerì una sintesi ancora maggiore. La nuova teoria è la miglior candidata per realizzare tale sintesi. Essa comprende i leptoni e i quark in una sola famiglia e fornisce i mezzi per trasformare un tipo di particella nell'altro. Nello stesso tempo le forze debole, forte ed elettromagnetica vengono trattate come aspetti di una sola forza fondamentale. Con una sola classe di particelle e una sola forza (oltre alla gravitazione), la teoria unificata è un modello di frugalità.

È noto che i leptoni e i quark hanno proprietà del tutto differenti; allora, come possono venire raggruppati in una sola famiglia? Le forze debole, forte ed elettromagnetica differiscono in intensità, in raggio di azione e in altre caratteristiche. Come si possono ricavare da una sola forza? La teoria unificata non tenta di nascondere tali differenze, ma afferma che non sono fondamentali. Le differenze sono sensibili principalmente perché oggi l'universo è molto freddo e quindi le particelle in generale hanno bassa energia. Se si potessero eseguire esperimenti a energia estremamente alta, l'unificazione diventerebbe chiara in tutta la sua semplicità. I leptoni e i quark si trasformerebbero liberamente gli uni negli altri e le tre forze avrebbero tutte la stessa intensità.

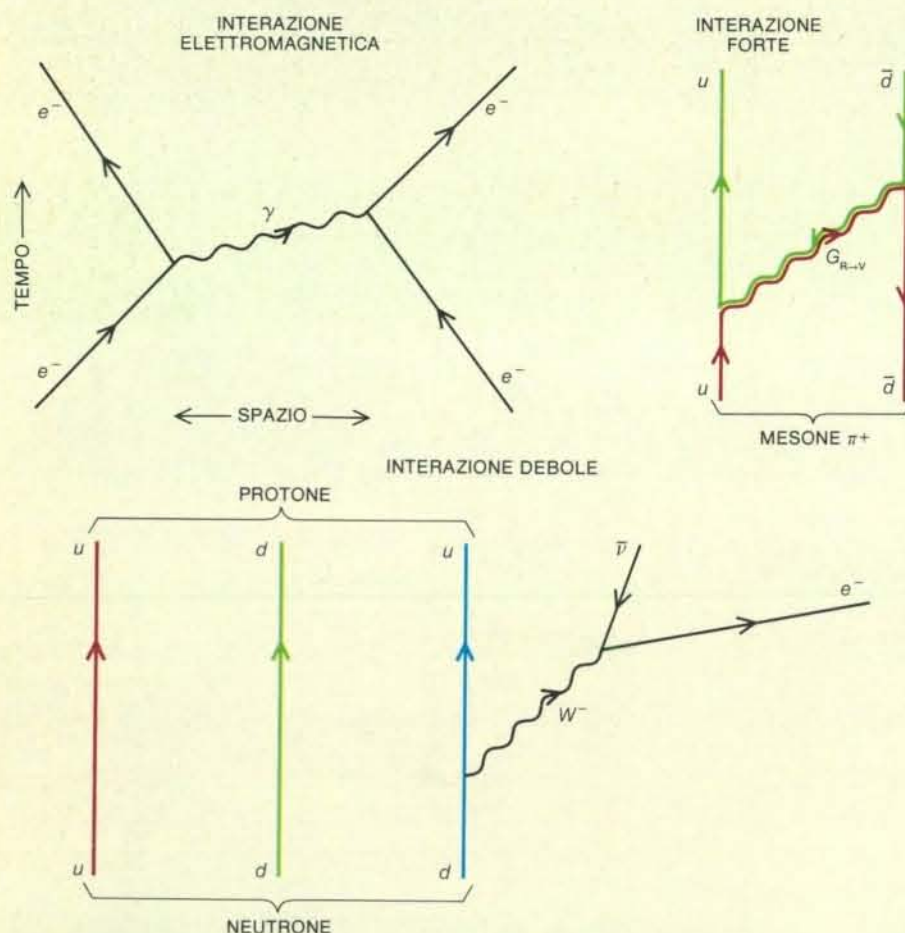
Si è calcolato che l'energia necessaria per poter vedere l'unificazione delle particelle e delle forze in questa forma drammatica è di circa 10^{15} gigaelettronvolt, o GeV. (Un GeV è l'energia impartita a un elettrone quando viene accelerato da una differenza di potenziale di un miliardo di volt.) Tale energia supera le possibilità perfino dei più grandi acceleratori di particelle in fase di progetto di un fattore di 10 bilioni, ed è estremamente improbabile che si possa mai raggiungere una tale energia in laboratorio. Si potrebbe perciò pensare che non si potrà mai verificare la teoria, ma non è così; la teoria ha conseguenze ben precise a energie facilmente accessibili.

In primo luogo, la teoria fornisce una spiegazione di numerosi aspetti ben noti del mondo fisico, che per lungo tempo erano parsi misteriosamente arbitrari. Riesce a spiegare la quantizzazione della carica elettrica: l'osservazione che la carica appare sempre in multipli discreti di una piccolissima carica fondamentale. Fornisce un valore delle intensità relative delle tre forze (misurate alle ordinarie energie di laboratorio) che è abbastanza bene in accordo con i risultati sperimentali. Potrebbe spiegare perché nell'universo

La simmetria cubica viene evidenziata quando si rappresentano graficamente certe proprietà delle particelle elementari. Le particelle sono membri delle famiglie chiamate leptoni e quark. La posizione di ogni particella nel piano orizzontale è determinata da tre tipi di «carica di colore». I quark chiamati *u* compaiono in tre colori e giacciono ai vertici di un triangolo equilatero; gli antiquark indicati con $\bar{d}$ hanno i tre corrispondenti anticolori e formano un triangolo con orientazione opposta. I leptoni, qui rappresentati dal positone (e^+) e dal neutrino (ν), non hanno carica di colore e si trovano al centro del piano. Quando si sposta una particella, verticalmente, di un tratto proporzionale alla sua carica elettrica, prende forma il cubo. Il fatto che questa disposizione delle particelle conduca a un solido semplice e simmetrico fa pensare all'esistenza di qualche collegamento tra i leptoni e i quark, collegamento che può essere spiegato da una teoria unificata.

c'è più materia che antimateria. È inoltre notevole il fatto che la teoria unificata preveda nuovi fenomeni che non si possono dedurre da precedenti teorie. La più rilevante di tali previsioni è il decadimento del protone, una particella che era stata ritenuta assolutamente stabile. Se il protone può decadere, lo stesso atomo è instabile e tutta la materia è mutevole.

La teoria unificata non è destinata a soppiantare le teorie già affermate delle forze debole, forte ed elettromagnetica; al contrario, le singole teorie sono comprese nella struttura di quella più estesa. Per questo, per poter spiegare la natura e l'origine della teoria unificata è meglio cominciare dalle singole teorie componenti, con le forze che le teorie descrivono e con le particelle elementari sulle quali le forze agiscono.

	LEPTONI		QUARK					
TERZA GENERAZIONE	ν_τ	0	*t*	+⅔	*t*	+⅔	*t*	+⅔
	τ^-	−1	*b*	−⅓	*b*	−⅓	*b*	−⅓
SECONDA GENERAZIONE	ν_μ	0	*c*	+⅔	*c*	+⅔	*c*	+⅔
	μ^-	−1	*s*	−⅓	*s*	−⅓	*s*	−⅓
PRIMA GENERAZIONE	ν_e	0	*u*	+⅔	*u*	+⅔	*u*	+⅔
	e^-	−1	*d*	−⅓	*d*	−⅓	*d*	−⅓

I leptoni e i quark differiscono per un certo numero di proprietà importanti e perciò sono stati solitamente classificati in famiglie distinte. Una delle differenze più rilevanti sta nella carica elettrica, qui mostrata per ogni particella: le cariche dei leptoni sono intere, mentre quelle dei quark sono frazionarie. Inoltre, i leptoni esistono come particelle libere, mentre i quark si trovano soltanto come costituenti delle particelle composte che vengono chiamate adroni. Si è soliti dividere i leptoni e i quark in tre generazioni; nella struttura della materia ordinaria hanno un posto solo le particelle appartenenti alla prima generazione. Il quark *t* non è stato osservato sperimentalmente.

Sono tre le forze della natura che producono le interazioni tra le particelle elementari. Ognuna di tali interazioni può essere descritta come lo scambio di una particella «virtuale», che è il vettore della forza. In un'interazione elettromagnetica le particelle dotate di carica elettrica scambiano un fotone (γ). La forza forte viene trasportata dai gluoni (G), scambiati da particelle con carica di colore. Le particelle con carica debole possono scambiare una W^- *(qui indicata)* o una W^+ o una Z^0. La carica di un'antiparticella è rappresentata qui da una freccia orientata all'indietro nel tempo.

Le evidenti differenze tra i leptoni e i quark sono sostanziali. Si conoscono sei leptoni, e l'elettrone può esserne considerato il prototipo. Esso ha una piccola massa, equivalente in unità di energia a circa 500 000 elettronvolt, e ha carica elettrica unitaria; per convenzione la carica dell'elettrone è negativa. Altri due leptoni, cioè il muone e la particella chiamata tau, hanno la stessa carica e paiono in realtà identici all'elettrone per tutte le proprietà eccettuata la massa. Il muone è oltre 200 volte più pesante dell'elettrone, mentre il leptone tau, che è stato scoperto solo cinque anni fa, ha una massa pari a quasi 3500 volte quella dell'elettrone.

Gli altri leptoni sono tre tipi di neutrini, che sono elettricamente neutri e hanno una massa piccolissima (ammesso che abbiano massa). Un neutrino è associato a ciascun leptone carico. Inoltre, per ognuno dei sei leptoni esiste un antileptone con la stessa massa, ma con carica elettrica opposta. Così l'antielettrone (o positone), l'antimuone e la particella antitau hanno tutti una carica + 1. Gli antineutrini, come gli stessi neutrini, non hanno carica elettrica.

Mentre i leptoni appaiono come particelle libere, nessuno è stato finora capace di osservare un quark isolato. I quark si osservano solo come costituenti delle particelle chiamate adroni, una grande e diversa classe che comprende il protone, il neutrone, il mesone pi e più di 100 altre particelle note.

Abbiamo indizi chiari dell'esistenza di cinque tipi di quark: essi vengono indicati con giù *(d)*, su *(u)*, strano *(s)*, incanto *(c)* e basso, o fondo *(b)*. Ne è stato previsto un sesto tipo chiamato alto, o sommità *(t)*, ma finora non è stato osservato. I tipi di quark vengono generalmente chiamati sapori; inoltre i quark hanno un'altra proprietà chiamata colore. (Sapore e colore sono etichette arbitrarie, che nulla hanno a che vedere con i sensi del gusto e della vista.) Ogni sapore di quark compare in tre colori: rosso, verde e blu. La proprietà del colore evidenzia un'ulteriore differenza tra i leptoni e i quark. I cinque o sei sapori di quark corrispondono all'incirca alle sei varietà di leptoni, ma non esiste una proprietà dei leptoni che corrisponda al colore dei quark. La distinzione ha conseguenze osservabili. La forza forte è un'interazione tra colori e, dato che i leptoni non hanno colore, non sono sensibili alla forza forte.

Un'altra proprietà distintiva dei quark è la loro carica elettrica. I quark *d*, *s* e *b* hanno una carica di − 1/3, mentre i quark *u*, *c* e *t* hanno una carica di + 2/3. Gli antiquark, indicati con $\bar{d}$, $\bar{u}$ e così via, hanno valori di carica elettrica opposti, quindi l'antiquark $\bar{d}$ ha una carica elettri-

L'elettromagnetismo viene descritto da una simmetria chiamata $U(1)$, termine preso a prestito dalla teoria matematica dei gruppi. $U(1)$ è il gruppo delle trasformazioni che possono essere applicate a un singolo oggetto o a una matrice uno per uno. Nella sua applicazione all'elettromagnetismo la simmetria $U(1)$ indica che la forza elettromagnetica non può modificare l'identità di una particella. La matrice uno per uno è occupata dal fotone, che può solo trasformare un elettrone in un altro elettrone.

ca di + 1/3, mentre l'antiquark $\bar{u}$ ha una carica di − 2/3. Gli antiquark hanno anche colori opposti, cioè antirosso, antiverde e antiblu.

I quark si possono combinare in due modi per formare un adrone. O si legano assieme tre quark, con un quark per ciascun colore, oppure si associa un quark di un dato colore a un antiquark del corrispondente anticolore. Si dice che tali combinazioni sono incolori e risulta che esse hanno un altro aspetto significativo. In tutte le combinazioni permesse la somma delle cariche frazionarie dei quark dà una carica intera; nessun'altra combinazione (tranne multipli di quelle permesse) ha tale proprietà. Il protone ha la composizione in quark *uud* con una carica elettrica totale di 2/3 + 2/3 − 1/3, ossia + 1.

	ROSSO	VERDE	BLU
ROSSO	$G_1 + G_2$	$G_{R\to V}$	$G_{R\to B}$
VERDE	$G_{V\to R}$	$G_1 + G_2$	$G_{V\to B}$
BLU	$G_{B\to R}$	$G_{B\to V}$	$G_1 + G_2$

La forza forte viene descritta da una teoria con una simmetria $SU(3)$, nella quale gli accoppiamenti dei gluoni con i quark si possono rappresentare con una matrice tre per tre. Qualsiasi colore dei quark nella colonna di sinistra della matrice può essere trasformato in uno qualsiasi degli altri colori nella riga superiore della matrice; la transizione viene mediata dal gluone indicato dall'intersezione tra la colonna e la riga. Un quark rosso può emettere un gluone $G_{R\to B}$ per diventare un quark blu. Due gluoni non cambiano colore, ma mediano trasformazioni come quella da rosso a rosso.

Il neutrone è formato dai quark *udd*, con cariche 2/3 − 1/3 − 1/3, con una carica complessiva nulla. Il mesone pi positivo è formato da un quark *u* e da un antiquark $\bar{d}$ e le cariche componenti di + 2/3 e di + 1/3 danno carica totale + 1.

Il fatto che tutti gli atomi siano elettricamente neutri implica che la carica del protone abbia esattamente lo stesso valore di quella dell'elettrone, anche se ovviamente i segni sono opposti. Per lo stesso motivo la carica del neutrone deve essere esattamente nulla. Da tali osservazioni deriva che le cariche dei quark devono essere esattamente commensurabili con quelle dei leptoni. Per esempio, la carica del quark *d* deve essere esattamente (e non approssimativamente) 1/3 di quella dell'elettrone. Tale relazione precisa tra particelle che appaiono indipendenti è un'altra proprietà apparentemente arbitraria che ci piacerebbe vedere spiegata in una teoria unificata.

Si è ormai soliti classificare i leptoni e i quark in tre generazioni, ognuna delle quali è costituita da un leptone carico, dal suo neutrino associato e da due quark, uno con carica − 1/3 e uno con carica + 2/3. La prima generazione comprende l'elettrone, il neutrino elettronico, il quark *d* e il quark *u*. Dato che ogni quark può avere tre colori, nella generazione ci sono otto particelle. Tutti gli atomi e tutta la materia ordinaria si possono ricavare da queste otto particelle; le generazioni di grado superiore si osservano quasi esclusivamente in esperimenti di laboratorio con particelle accelerate. Nella teoria unificata le tre generazioni sono descritte indipendentemente, ma praticamente nello stesso modo. Parlerò pertanto solo della prima generazione.

Delle tre forze che qui esaminerò, l'elettromagnetismo è stata la prima a essere descritta da una teoria precisa; in realtà, la precisione della teoria non è mai stata superata. La teoria è l'elettrodinamica quantistica, o QED, e si è sviluppata in un periodo di circa 25 anni, culminato all'inizio degli anni cinquanta. Essa è servita da modello per teorie relative alle altre forze.

Il concetto di forza è strettamente correlato con quello di carica. La carica elettrica è la proprietà attribuita a una particella sensibile alla forza elettromagnetica e l'entità della carica determina la risposta. Quando due particelle cariche si avvicinano una all'altra, si produce un'attrazione o una repulsione il cui valore assoluto è direttamente proporzionale al prodotto delle cariche. La forza è inoltre inversamente proporzionale al quadrato della distanza tra le cariche. Queste due regole costituiscono la legge di Coulomb della forza elettrica. È importante notare che se una delle particelle ha carica nulla, non c'è attrazione né repulsione; tali particelle neutre non sono direttamente sensibili alla forza elettromagnetica.

Che intensità ha l'interazione elettromagnetica tra particelle cariche? Per qualsiasi coppia di particelle la risposta dipende dalle cariche e dalla loro distanza, ma la legge di Coulomb può dare una risposta assoluta. Supponiamo di moltiplicare la forza tra le due particelle per il quadrato della loro distanza; il prodotto è una misura dell'intensità dell'interazione elettromagnetica che è indipendente dalla distanza tra le particelle, anche se dipende dal sistema di unità in cui viene espressa la distanza. Dividendo la velocità della luce per la costante di Planck (due grandezze intrinseche alla struttura del mondo relativistico quantomeccanico) si ha un risultato indipendente dalle unità. Il risultato è un numero puro, ossia adimensionale, e ha lo stesso valore se le misure sono in grammi, centimetri e secondi o in tonnellate, metri e anni, con la sola condizione che la velocità della luce e la costante di Planck siano espresse nelle stesse unità impiegate per le misure.

L'intensità di una qualsiasi interazione elettromagnetica dipende evidentemente ancora dall'entità delle cariche che la subiscono. Così l'interazione sarà quattro volte più intensa se si raddoppiano le cariche. Tuttavia, dato che la carica elettrica è quantizzata, l'interazione di due protoni o di due elettroni riveste un ruolo particolare. Tutte le particelle osservate singolarmente (cioè, tutte tranne i quark) hanno cariche multiple intere della carica del protone e quindi l'interazione protone-protone è una misura della intensità minima della forza elettromagnetica. Questa quantità è detta costante di accoppiamento elettromagnetica e costituisce una misura assoluta dell'intensità dell'interazione. Le determinazioni sperimentali della costante di accoppiamento danno un valore di circa 1/137. Essendo tale valore minore di 1, l'interazione elettromagnetica è decisamente debole.

È opportuno notare che la quantizzazione della carica non è richiesta né prevista dall'elettrodinamica quantistica, ma è un fatto sperimentale. La teoria andrebbe altrettanto bene se ci fossero particelle osservabili con cariche frazionarie o perfino con quantità di carica irrazionali, quali π o la radice quadrata di 2.

Nell'elettrodinamica quantistica l'interazione di due particelle cariche, quali due elettroni, è correlata allo scambio di una terza particella. La particella intermedia è un fotone, un quanto di radiazione elettromagnetica. Il fotone è una particella priva di massa non dotata di carica elettrica propria e in moto (per definizione) con la stessa velocità della luce. La descrizione della forza elettromagnetica come uno scambio di fotoni evita la problematica nozione di azione a distanza. L'interazione è confinata a due eventi puntiformi: l'emissione e l'assorbimento del fotone. La descrizione introduce però al tempo stesso un altro problema ugualmente serio: lo scambio del fotone pare violare le leggi della natura che richiedono la conservazione dell'energia e della quantità di moto.

La violazione può essere illustrata supponendo due elettroni tenuti in quiete a una piccola distanza. Dato che tra gli elettroni si potrebbe misurare una forza, si deve supporre che avvenga uno scambio di fotoni. Solitamente, quando viene emesso un fotone, esso porta via una parte dell'energia e della quantità di moto della particella emittente; allo stesso modo, quando viene assorbito un fotone, esso apporta energia e quantità di moto alla particella assorbente. In questo modo si conservano l'energia e la quantità di moto totali del sistema. Nella situazione in esame, tuttavia, la particella emittente è mantenuta in quiete, quindi la sua energia e la sua quantità di moto non possono cambiare e altrettanto accade per la particella assorbente. Evidentemente, il fotone scambiato gode di proprietà alquanto particolari, diverse da quelle dei fotoni che formano la luce solare o le radioonde. Per tali differenze il fotone scambiato è detto fotone virtuale.

La spiegazione delle peculiari proprietà del fotone virtuale risiede nel principio di indeterminazione introdotto in meccanica quantistica da Werner Heisenberg. Il principio di indeterminazione non invalida i principi di conservazione dell'energia e della quantità di moto, ma consente che una violazione di tali principi passi inosservata se rettificata con sufficiente prontezza. Gli elettroni in quiete hanno la stessa energia e la stessa quantità di moto prima e dopo l'emissione e l'assorbimento del fotone virtuale; i principi di conservazione appaiono violati solo durante il breve passaggio del fotone. Il principio di indeterminazione afferma che tale violazione può essere tollerata se non dura troppo a lungo o se non si estende a distanza troppo grande.

Ma che cos'è «troppo» a lungo e che cosa vuol dire «troppo» distante? Le risposte dipendono dall'entità della violazione apparente: quanto maggiore è lo squilibrio, nell'energia e nella quantità di moto, causato dall'emissione del fotone virtuale, tanto più rapidamente il fotone deve essere riassorbito. Ad alta energia un fotone virtuale può sopravvivere per poco tempo, mentre un fotone a bassa energia gode di un lungo periodo felice prima che si ristabilisca l'equilibrio. In altri termini, il prodotto dello squilibrio energetico per la vita media del fotone non può essere maggiore della costante di Planck. La minima energia che può avere qualsiasi particella è l'energia equivalente alla massa di riposo della particella e quindi il massimo raggio d'azione di una particella virtuale è inversamente proporzionale alla sua massa. Il raggio d'azione della forza elettromagnetica pare sia infinito, perciò si pensa che la massa di riposo del fotone sia esattamente nulla.

La presenza di particelle virtuali complica enormemente la struttura dell'universo. A causa delle particelle virtuali il vuoto non è uno spazio semplicemente privo di materia. Un fotone virtuale può comparire spontaneamente in un qualsiasi istante per scomparire di nuovo nell'intervallo di tempo consentito dal principio di indeterminazione. Allo stesso modo possono essere create altre particelle virtuali, comprese le particelle elettricamente cariche; l'unica condizione è che le par-

CARICHE DI COLORE

		R — V	V — B	B — R	
QUARK	ROSSO	+½	0	−½	=0
	VERDE	−½	+½	0	=0
	BLU	0	−½	+½	=0
		=0	=0	=0	
GLUONI	G_1	0	0	0	=0
	G_2	0	0	0	=0
	$G_{R\to V}$	+1	−½	−½	=0
	$G_{V\to R}$	−1	+½	+½	=0
	$G_{V\to B}$	−½	+1	−½	=0
	$G_{B\to V}$	+½	−1	+½	=0
	$G_{R\to B}$	+½	+½	−1	=0
	$G_{B\to R}$	−½	−½	+1	=0
		=0	=0	=0	

Le cariche di colore dei quark e dei gluoni si possono identificare come rosso meno verde (R − V), verde meno blu (V − B) e blu meno rosso (B − R). Ciascuno dei colori dei quark rosso, verde e blu viene definito da una combinazione delle tre cariche. È significativo che le cariche che contribuiscono a ciascun colore abbiano per somma zero. Questo schema implica che le tre cariche non siano del tutto indipendenti e in realtà due qualsiasi di esse bastano a identificare il colore. (Qui sono indicate tutte e tre le cariche per chiarezza.) Per un tripletto di quark, un quark per ciascun colore, le somme dei valori di ciascuna carica sono ancora nulle. Sei dei gluoni hanno cariche di colore, con i valori necessari per trasformare un quark da un colore a un altro. Lo schema delle cariche nel tripletto dei quark e la presenza di cariche nei gluoni impongono la quantizzazione della carica di colore, i cui soli valori possibili sono multipli discreti di 1/2 unità.

Il procedimento di «montaggio» degli antiquark e dei gluoni prevede le loro proprietà di colore supponendo che siano combinazioni di quark tratti dal tripletto fondamentale di colore. Un antiquark qualsiasi (che deve avere cariche di colore opposte a quelle del quark corrispondente) si può «costruire» sommando le cariche di colore di due quark. Un qualsiasi gluone si può ricavare con i colori di un quark e di un antiquark, o, dato che lo stesso antiquark può essere decomposto, dai colori di tre quark. Appaiono giustificate tutte le possibili combinazioni. Il procedimento è solo formale; non si dovrebbero considerare antiquark e gluoni come composti fisici di quark.

ticelle con carica elettrica devono comparire e scomparire in coppie associate di particella e antiparticella. Questo processo ha profonde conseguenze sulla teoria dell'elettromagnetismo.

Si consideri ciò che accade quando un elettrone reale è immerso in una nube di fotoni virtuali e di coppie virtuali elettrone-positone. L'influenza dei fotoni è modesta, ma le particelle cariche virtuali si polarizzano: le cariche negative virtuali vengono respinte dalle cariche negative reali, mentre le cariche positive virtuali vengono attratte dall'elettrone reale. Il risultato è che l'elettrone si circonda a breve distanza di una nube di cariche positive, che schermano in parte la carica dell'elettrone.

Da questa analisi deriva che la carica «nuda» dell'elettrone è molto maggiore di quella misurata. In pratica, in elettrodinamica quantistica si assume che la carica nuda sia infinita. La carica che si misura è semplicemente il residuo finito che rimane quando si sottrae la carica di schermo dalla carica nuda. Se si potesse misurare la carica dell'elettrone a brevissima distanza, si scoprirebbe che essa aumenta entrando nello strato di schermo. Un'ulteriore conseguenza è che la costante di accoppiamento dell'elettromagnetismo non è costante affatto, ma varia con la distanza alla quale le particelle cariche interagiscono una con l'altra. La costante di accoppiamento aumenta (cioè l'interazione elettromagnetica diventa più intensa) al diminuire del raggio d'azione. La costante di accoppiamento misurata di circa 1/137 è osservata a distanze atomiche di circa 10^{-8} centimetri.

Pur nell'effimero campo delle particelle virtuali c'è un principio di conservazione che non viene mai violato: la conservazione della carica elettrica. Dato che lo stesso fotone è neutro, la carica viene automaticamente conservata nello scambio di un fotone virtuale; nessuna carica viene alterata. Inoltre, quando la materia carica si crea o si annichila, lo fa sempre a coppie di particelle e di antiparticelle, in modo che la somma delle cariche prima e dopo l'evento sia sempre la stessa.

La conservazione della carica elettrica e l'assenza di massa del fotone sono correlate a un gruppo di simmetrie nel sistema matematico che descrive l'elettrodinamica quantistica. Il gruppo di simmetrie è designato con $U(1)$ e quindi la QED è una teoria $U(1)$. $U(1)$ è un termine impiegato nella teoria matematica dei gruppi. L'1 si riferisce al fatto che il fotone interagisce soltanto con un tipo di particelle alla volta. Il fotone non trasforma mai un tipo di particelle in un altro tipo. La forza forte e la forza debole sono più complicate da questo punto di vista e i gruppi che le descrivono sono più complessi.

La teoria prevalente della forza forte è modellata direttamente sulla elettrodinamica quantistica ed è chiamata cromodinamica quantistica, o QCD, dove «cromo» sta a indicare che le forze non agiscono tra cariche elettriche, ma tra cariche di colore. Come nella QED, il valore assoluto della forza tra due cariche è proporzionale al prodotto delle cariche. Le particelle che non hanno carica di colore non sono soggette alla forza. Una costante di accoppiamento adimensionale definisce l'intensità intrinseca dell'interazione. La costante di accoppiamento è maggiore della costante dell'elettromagnetismo, come del resto si può prevedere per una forza che viene chiamata forte.

La QCD, pur se costruita sugli stessi principi della QED, è una teoria più elaborata. La causa principale dell'ulteriore complessità è la molteplicità delle cariche di colore. Mentre l'elettromagnetismo è associato a un solo tipo di carica, la forza forte agisce su tre colori: rosso, verde e blu. Ciascuno rappresenta una combinazione di cariche di colore fondamentali.

Vi sono molti modi di analizzare le cariche di colore. Quello che adotterò parte ammettendo l'esistenza di tre tipi di cariche di colore. Chiamerò le cariche rosso meno verde (R – V), verde meno blu (V – B) e blu meno rosso (B – R). Ogni carica può avere un valore + 1/2, – 1/2 o 0, e ogni colore di quark è contraddistinto da una particolare combinazione di valori. Un quark è rosso se possiede una carica R – V di + 1/2, una carica V – B nulla e una carica B – R di –1/2. Un quark verde ha cariche di colore R – V = – 1/2, V – B = = + 1/2 e B – R = 0. In un quark blu le tre cariche sono R – V = 0, V – B = = – 1/2 e B – R = + 1/2. Gli anticolori associati agli antiquark si ottengono semplicemente invertendo i segni di tutte le cariche.

Su questa distribuzione di cariche si possono fare diverse osservazioni. In primo luogo, esistono 27 combinazioni possibili delle tre cariche, quando ogni carica può avere uno qualsiasi dei tre valori. Ciononostante, pare che in natura esistano soltanto i quark con le tre combinazioni che danno i colori rosso, verde e blu. In secondo luogo, questo sottogruppo delle possibili cariche di colore è altamente caratteristico. Ognuna delle combinazioni osservate è tale che la somma delle tre cariche è nulla e le combinazioni osservate sono le sole che godono di tale proprietà. (Esiste in realtà un'altra combinazione con una carica di colore totale nulla: la combinazione nella quale ciascuna carica è nulla. Una particella che non ha alcuna carica di colore, però, non è un quark.)

La scoperta che la somma delle tre cariche di colore è sempre nulla indica che una delle tre cariche non è indipendente dalle altre due. Se si conoscono due qualsiasi delle tre cariche, la terza si può calcolare per differenza. Si può quindi concludere che esistono in pratica solo due varietà di cariche di colore, che sono sufficienti per specificare completamente i tre colori. Non importa quali sono le due cariche che si considerano fondamentali e quale venga scartata; qui supporrò che le cariche R – V e V –B siano fondamentali, ma manterrò, per chiarezza, la carica B – R, anche se le informazioni che essa fornisce sono ridondanti.

Si può notare un'ulteriore relazione tra le cariche. In uno stato composto da un quark rosso, uno verde e uno blu, la quantità totale di ogni carica di colore è ancora nulla. In altri termini, la combinazione dei tre colori origina uno stato di colore neutro, allo stesso modo in cui la combinazione di un elettrone e di un protone crea uno stato (l'atomo di idrogeno) neutro rispetto alla carica elettrica. È proprio in questo modo che si formano gli adroni di colore neutro quale il protone. Si crea un sistema incolore anche combinando un colore con il corrispondente anticolore; le cariche di colore, essendo opposte, si annullano esattamente. L'associazione di un colore al suo anticolore è l'altra formula per formare un adrone, sul modello del mesone pi. Eccettuati i multipli di queste combinazioni (quale un sistema a sei quark che comprende due quark per ogni colore), non esiste altro modo di combinare quark colorati e quindi tutte le cariche di colore hanno una somma nulla.

Il meccanismo con cui viene trasmessa la forza forte è confrontabile al corrispondente meccanismo dell'elettromagnetismo: l'interazione tra due particelle cariche viene descritta come lo scambio di una terza particella. Però, ancora una volta, la cromodinamica quantistica è una teoria più elaborata. Mentre la QED ha un solo fotone privo di massa, la QCD ha otto particelle prive di massa chiamate gluoni. Inoltre, mentre il fotone non ha carica elettrica, alcuni gluoni trasportano una carica di colore. La presenza di particelle portatrici di carica modifica fondamentalmente il carattere della forza.

I gluoni, essendo essi stessi carichi, non solo possono trasmettere la forza forte, ma possono anche alterare i colori dei quark. Al contrario, l'emissione o l'assorbimento di fotoni non può mai alterare la carica elettrica di una particella. Esistono nove transizioni possibili tra i colori dei quark, definite da una matrice tre per tre. Per esempio, un quark rosso può essere trasformato in un quark rosso (trasformazione identità), in un quark verde o in quark blu. Le tre trasformazioni identità (il rosso diventa rosso, il verde verde e il blu blu) costituiscono gli elementi della diagonale della matrice. Ne consegue che i gluoni responsabili delle transizioni di identità non possono avere carica di colore, altrimenti modificherebbero i colori dei quark. Si potrebbe concludere che debbano esistere tre di tali gluoni incolori, ciascuno per ogni trasformazione identità. Tuttavia, dal momento che sono necessarie due cariche di colore indipendenti per specificare i tre colori del quark, esistono soltanto due gluoni incolori. Li chiamerò G_1 e G_2.

Le sei restanti transizioni tra i colori dei quark comportano variazioni di colore. Ognuna di queste transizioni è associata al proprio gluone e ognuno di questi sei gluoni porta una carica di colore. Indicherò i gluoni con carica di colore in modo descrittivo. Per esempio, un gluone dal rosso al verde, o $G_{R \to V}$, può essere emesso da un quark rosso, che viene di con-

seguenza trasformato in un quark verde.

Le cariche di colore trasportate dai gluoni si possono ricavare in base al requisito della conservazione della carica di colore. Si consideri il processo nel quale un quark passa da rosso a verde emettendo un gluone $G_{R \to V}$. Nel corso della transizione la carica R – V del quark passa da + 1/2 a – 1/2; se la quantità di carica totale deve restare invariata il gluone deve avere una carica R – V pari a + 1. Allo stesso modo la carica V – B del quark passa da 0 a + 1/2, e il gluone deve portar via una carica V – B di – 1/2. La carica B – R del quark si trasforma da – 1/2 a 0, che implica nuovamente una carica di –1/2 di un gluone B – R. Quindi le cariche di colore del gluone sono rispettivamente + 1, – 1/2 e – 1/2. Il quark che media la trasformazione inversa, da verde a rosso, deve avere cariche di ugual valore assoluto, ma di segno opposto.

La presenza delle cariche di colore del gluone ha un'ulteriore conseguenza: garantisce automaticamente la quantizzazione della carica di colore. In elettromagnetismo, in linea di principio, un fotone potrebbe essere emesso o assorbito da una particella con carica elettrica qualsiasi. Le particelle con carica di colore però possono interagire scambiando gluoni soltanto se le cariche differiscono per multipli di 1/2. Si può anche dimostrare che le cariche di colore del sistema devono essere simmetriche rispetto a zero, cioè la somma di tutte le cariche positive e la somma di tutte le cariche negative devono essere uguali in valore assoluto.

La quantizzazione della carica di colore si può dimostrare in un altro modo. Qualsiasi sistema di particelle dotate di colore si può costruire a partire dal seguente semplicissimo sistema: il tripletto costituito da un quark rosso, uno verde e uno blu. Il tripletto di antiquark si può formare combinando i quark a coppie. Non intendo affermare che un antiquark fisico sia formato da due quark in uno stato legato. Ciononostante, tale sintesi fornisce correttamente tutte le proprietà di colore degli antiquark. Si noti che un quark rosso ha cariche di colore R – V, V – B e B – R rispettivamente uguali a + 1/2, 0 e –1/2. Un antiquark antirosso deve avere cariche opposte: – 1/2, 0 e + 1/2. Sono proprio questi i valori trovati sommando le cariche di un quark verde (– 1/2, + 1/2 e 0) e di un quark blu (0, – 1/2 e + 1/2). Quindi l'antirosso è in un certo senso equivalente alla somma del verde e del blu. Allo stesso modo l'antiverde è formato dal rosso più il blu e l'antiblu dal rosso più il verde. Questa notevole corrispondenza è una conseguenza del modo in cui sono distribuite le cariche nel tripletto dei quark. Essendo nulla la carica totale del tripletto, la somma di due cariche qualsiasi deve uguagliare l'opposto della restante carica.

I gluoni si possono costruire in modo analogo da un quark e da un antiquark, benché anche qui non si possa dedurre alcun processo di costruzione fisico. Il gluone da rosso a verde, con cariche + 1, – 1/2 e – 1/2, si può pensare composto da un quark rosso (+ 1/2, 0 e – 1/2) e da un antiquark antiverde (+ 1/2, – 1/2 e 0). Naturalmente, l'antiquark antiverde si può ulteriormente decomporre in un quark rosso e in un quark blu e quindi il gluone da rosso a verde ha le proprietà di colore di due quark rossi e di uno blu.

	CARICA DEBOLE	CARICA $U(1)$	CARICA ELETTRICA	PARTICELLE	TRANSIZIONI
DOPPIETTI	+½	–½	0	$\nu_{SINISTRA}$	W^+ ↓ ↑ W^-
	–½		–1	$e^-_{SINISTRA}$	
	+½	+⅙	+⅔	$u_{SINISTRA}$	W^+ ↓ ↑ W^-
	–½		–⅓	$d_{SINISTRA}$	
	+½	+½	+1	e^+_{DESTRA}	W^+ ↓ ↑ W^-
	–½		0	$\bar{\nu}_{DESTRA}$	
	+½	–⅙	+⅓	$\bar{d}_{DESTRA}$	W^+ ↓ ↑ W^-
	–½		–⅔	$\bar{u}_{DESTRA}$	
SINGOLETTI	0	–1	–1	e^-_{DESTRA}	
	0	+1	+1	$e^+_{SINISTRA}$	
	0	+⅔	+⅔	u_{DESTRA}	
	0	–⅔	–⅔	$\bar{u}_{SINISTRA}$	
	0	–⅓	–⅓	d_{DESTRA}	
	0	+⅓	+⅓	$\bar{d}_{SINISTRA}$	

La carica debole dipende dall'orientamento di una particella ed è anche curiosamente correlata alla carica elettrica. Le particelle sinistrorse e le antiparticelle destrorse formano doppietti nelle interazioni deboli e a esse vengono assegnate cariche deboli di più o meno 1/2. La W^+ e la W^-, che mediano la forza debole, trasformano un membro di un doppietto nell'altro membro. Le particelle destrorse e le antiparticelle sinistrorse restano singoletti e non hanno carica debole, e tra esse non è di conseguenza possibile alcuna transizione debole. La carica elettrica di ogni particella consiste nella somma della carica debole e di un'altra grandezza, detta carica $U(1)$, che è uguale alla carica elettrica media delle particelle del singoletto o del doppietto. Questa relazione presente tra le cariche indica l'esistenza di una connessione fondamentale tra la forza debole e l'elettromagnetismo.

C'è tuttavia un'altra conseguenza dell'assegnazione di cariche di colore ai gluoni. Come detto sopra, un elettrone nel vuoto si circonda di una nube di fotoni virtuali e di coppie elettrone-positone virtuali; le particelle virtuali cariche si polarizzano e schermano una parte della carica nuda dell'elettrone. Con lo stesso meccanismo un quark nel vuoto viene racchiuso in una nube di gluoni virtuali e di coppie quark-antiquark virtuali ma il risultato è del tutto differente. La nube di quark e antiquark virtuali si polarizza nel solito modo, con gli antiquark addensati nei pressi della carica di colore reale e tendenti a schermarla. Invece, i gluoni virtuali hanno un effetto opposto. La carica di colore predominante dei gluoni è la stessa della carica del quark. Inoltre, i gluoni virtuali sono più numerosi dei quark virtuali e quindi l'influsso dei gluoni è più intenso. Il risultato è che la carica del quark appare diffusa nello spazio e la carica efficace diminuisce quanto più ci si avvicina al quark.

In assenza di cariche del gluone ci si potrebbe aspettare che la forza forte vari con la distanza come l'elettromagneti-

smo. Essendo i gluoni, come il fotone, privi di massa, la forza dovrebbe avere un raggio d'azione infinito, ma dovrebbe diminuire di intensità con il quadrato della distanza. Il fatto che i gluoni trasportino cariche di colore altera il carattere della forza. Dato che la nube di gluoni virtuali diffonde esternamente la carica di colore, la forza di colore tra due quark non aumenta così rapidamente come quella elettromagnetica quando diminuisce la distanza tra le particelle. Il risultato è che la costante di accoppiamento della QCD diminuisce al diminuire della distanza alla quale viene misurata (invece la costante di accoppiamento della QED aumenta a distanze minime). Si dice che i quark sono asintoticamente liberi, cioè che la costante di accoppiamento della QCD si annulla quando la distanza tende a zero.

La libertà asintotica venne scoperta da H. David Politzer, che ora lavora al California Institute of Technology, e da David Gross e Franck Wilczek della Princeton University. È stata verificata e confermata in un certo numero di esperimenti che sondano la struttura a quark degli adroni a piccole distanze. La natura delle interazioni forti tra i quark a distanze maggiori non è ancora ben definita, ma sembra che la forza non possa diminuire con il quadrato della distanza e possa invece restare costante indipendentemente dalla distanza. Sarebbe allora necessaria una quantità illimitata di energia per separare due cariche di colore, e ciò spiegherebbe perché i quark appaiono permanentemente confinati negli adroni.

Si dice che la cromodinamica quantistica è una teoria $SU(3)$, un altro termine della teoria dei gruppi. Il 3 si riferisce ai tre colori che vengono trasformati uno nell'altro dai gluoni. S indica che la somma delle cariche di colore in ogni famiglia $SU(3)$ è nulla. In stretta analogia con la $U(1)$ della QED, la $SU(3)$ della QCD descrive un gruppo di simmetrie della teoria associato alla conservazione della carica di colore e all'assenza di massa dei gluoni.

Il più elevato ordine di simmetria della teoria $SU(3)$ del colore si può intuire con una presentazione geometrica. Le tre cariche di colore R – V, V – B e B – R si possono rappresentare con la posizione lungo tre assi complanari. Gli assi sono disposti simmetricamente ad angoli di 120 gradi uno dall'altro. Se si rappresentano in tale grafico i tre colori secondo le loro cariche di colore componenti, si trova che essi giacciono ai vertici di un triangolo equilatero. Gli anticolori sono opposti ai corrispondenti colori e formano perciò un altro triangolo ruotato di 180 gradi rispetto al primo. I due triangoli sovrapposti formano una stella di Davide.

Un'idea della ancor più grande simmetria della teoria unificata si può avere aggiungendo al grafico una terza dimensione. Supponiamo che le cariche di colore già rappresentate nel grafico siano quelle del quark u e dell'antiquark $\bar{d}$. Aggiungiamo ora due leptoni: il neutrino elettronico e il positone. Dal momento che i leptoni non hanno carica di colore, essi si trovano all'origine dei tre assi al centro del piano. La terza dimensione è la carica elettrica: ogni particella deve essere spostata verticalmente di una quantità proporzionale alla sua carica elettrica. Il neutrino resta al suo posto, ma i tre antiquark $\bar{d}$ vengono spostati verso l'alto di un terzo di unità, i quark u si spostano verso l'altó di due terzi di unità e il positone si sposta verso l'alto di un'unità. Se si scelgono opportunamente le scale verticale e orizzontale, le otto particelle definiscono i vertici di un cubo appoggiato su uno spigolo. Il fatto che quark e leptoni possano essere disposti nella configurazione di questo semplice solido suggerisce qualche correlazione fondamentale tra essi.

Per poter discutere della forza debole, è necessario introdurre un'altra proprietà delle particelle elementari: il momento angolare di spin. Si è osservato che tutti i leptoni e tutti i quark hanno la stessa quantità fissa di momento angolare, uguale a 1/2 se misurata in unità fondamentali. Si possono pensare le particelle come ruotanti (come la Terra o come una trottola) attorno a un asse interno, ma senza perdere energia. Il momento angolare è rappresentato da un vettore, o freccia, lungo l'asse di spin.

Una particella con mezzo quanto di spin intrinseco può avere soltanto due orientazioni possibili; nel caso più semplice, quando la particella è in moto, il vettore di spin può puntare nella stessa direzione di moto della particella o in direzione opposta. Le due orientazioni rappresentano due stati distinti della particella. Lo stato in cui il vettore è parallelo alla direzione di moto è detto destrorso, perché quando le dita della mano destra vengono avvolte attorno alla particella nello stesso verso dello spin, il pollice indica la direzione di moto. Quando l'asse di spin ha direzione opposta, è il pollice della mano sinistra a indicare la direzione di moto, e la particella è detta sinistrorsa.

In generale l'orientamento di una particella può essere invertito semplicemente arrestando la particella e accelerandola in direzione opposta senza perturbarne lo spin. Quindi la maggior parte delle particelle necessariamente ha componenti sia sinistrorse che destrorse. Fanno eccezione le particelle prive di massa e la ragione per cui esse costituiscono eccezioni è dovuta al fatto che una particella priva di massa deve sempre muoversi con la velocità della luce e non può mai essere portata allo stato di quiete. L'orientamento di una particella priva di massa perciò non può mai cambiare. Tra i quark e i leptoni le sole particelle che potrebbero essere prive di massa sono i neutrini. Sperimentalmente sono stati osservati soltanto neutrini sinistrorsi e antineutrini destrorsi; si pensa che i neutrini destrorsi e gli antineutrini sinistrorsi non esistano.

L'introduzione dell'orientamento ha l'effetto di raddoppiare quasi il numero di particelle elementari distinguibili, un numero già abbastanza grande. Nella prima generazione di particelle ci sono due leptoni (l'elettrone e il neutrino elettronico) e due sapori di quark (u e d). I tre colori dei quark danno un totale di otto particelle, che, unite alle otto antiparticelle corrispondenti, raggiungono un totale di 16. Se tutte le particelle avessero sia componenti sinistrorse che destrorse, l'inclusione dell'orientamento raddoppierebbe ulteriormente il numero. In effetti, non essendovi neutrini destrorsi né antineutrini sinistrorsi, il numero totale di stati di particelle e di antiparticelle distinti è 30. Sono questi 30 stati a dover essere sistemati in una teoria unificata.

Si devono distinguere gli stati di diverso orientamento, perché la forza debole agisce diversamente sulle componenti sinistrorsa e destrorsa di una particella. Come le altre forze, la forza debole è associata a una carica e l'intensità intrinseca dell'interazione debole può essere definita mediante una costante di accoppiamento adimensionale. La carica debole è però insolita: viene assegnata in base all'orientamento. Soltanto le particelle sinistrorse e le antiparticelle destrorse trasportano una carica debole; le particelle destrorse e le antiparticelle sinistrorse sono insensibili alla forza debole e non partecipano alle interazioni deboli.

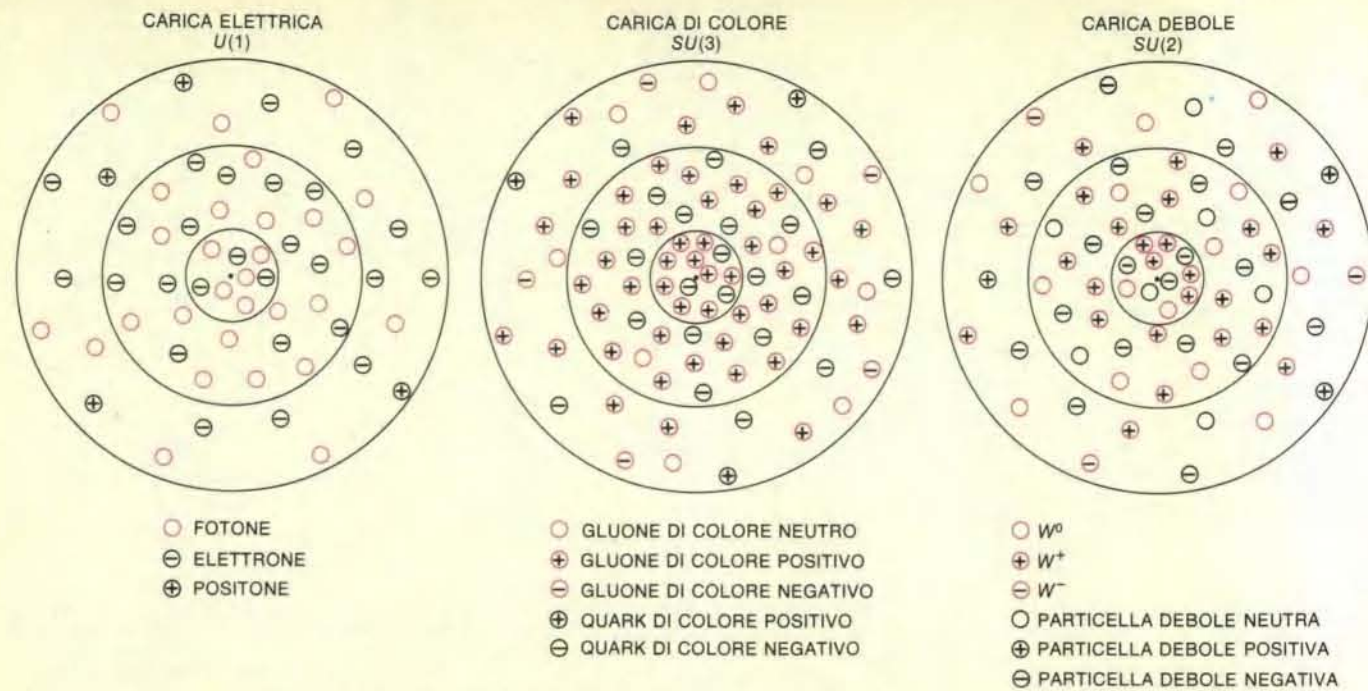
Poiché le cariche deboli dell'elettrone sinistrorso e destrorso (per fare un esempio) sono diverse, la carica debole non si può conservare. Il valore della carica debole dipende dal modo in cui si muove l'elettrone e il valore deve cambiare quando cambia il moto. La carica debole potrebbe conservarsi soltanto se i leptoni e i quark fossero tutti privi di massa, dato che in tal caso nessuna delle particelle potrebbe arrestarsi e invertire direzione.

La forza debole agisce su doppietti di particelle. La teoria che la descrive è una teoria $SU(2)$, nella quale i due membri di un doppietto possono trasformarsi uno nell'altro. Per esempio, il neutrino sinistrorso e l'elettrone sinistrorso possono formare un doppietto; a essi vengono assegnate cariche deboli rispettivamente di + 1/2 e – 1/2. Il quark sinistrorso u e il quark sinistrorso d formano un secondo doppietto (o tre doppietti se si conta separatamente ogni colore) e hanno anch'essi rispettivamente cariche deboli + 1/2 e – 1/2. I restanti doppietti sono formati da quattro antiparticelle destrorse: il positone, l'antineutrino elettronico, l'antiquark $\bar{d}$ e l'antiquark $\bar{u}$. Ogni antiparticella destrorsa ha una carica debole opposta a quella della corrispondente particella sinistrorsa. Restano da giustificare sei particelle: le componenti destrorse dell'elettrone, del quark d e del quark u e le componenti sinistrorse del positone, dell'antiquark $\bar{d}$ e dell'antiquark $\bar{u}$. Esse non formano doppietti, ma restano isolate come singoletti e hanno una carica debole uguale a zero.

Tre particelle associate alla simmetria debole $SU(2)$ mediano le transizioni tra i membri di ciascun doppietto. Le particelle intermedie sono la W^+, con carica debole e carica elettrica entrambe di + 1, la W^-, con cariche debole ed elettrica –1 e la W^0, neutra sia rispetto alla forza debole sia rispetto a quella elettromagnetica. La W^0, come il fotone e i gluoni G_1 e G_2, trasporta una forza tra particelle cariche, ma non altera alcuna delle loro proprietà. Invece, la W^+ e la W^- trasformano i sapori delle particelle. Un elettrone sinistrorso può emettere una W^- e trasformarsi quindi in un neutrino sinistrorso; nel processo la carica elettrica cambia da –1 a 0 e la carica debole da – 1/2 a + 1/2. Il più comune dei processi deboli è il decadimento nucleare beta, nel quale un neutrone *(udd)* emette un elettrone e un antineutrino trasformandosi in un protone *(uud)*. Analizzati su una scala più dettagliata l'evento inizia quando un quark d emette una W^- virtuale e diventa un quark u; la W^- decade successivamente in un elettrone e in un antineutrino.

La forza debole e l'elettromagnetismo si possono descrivere congiuntamente con una teoria dotata di una simmetria rappresentata dal prodotto di due gruppi: $SU(2) \times U(1)$. La parte $SU(2)$ dell'interazione induce tutte le possibili trasformazioni di due oggetti o di una matrice due per due. Gli oggetti sono i membri dei doppietti deboli, rappresentati qui dalle componenti sinistrorse dell'elettrone e del neutrino. Esistono tre particelle $SU(2)$ e tutte e tre le particelle sono prive di massa: la W^+ e la W^- trasformano un elettrone in un neutrino e viceversa, mentre la W^0 media le operazioni di identità elettrone-elettrone e neutrino-neutrino. La parte $U(1)$ dell'interazione è associata a un'altra particella intermediaria, la V^0, la quale è capace soltanto di trasformazioni identità. Nella teoria combinata $SU(2) \times U(1)$ la W^0 e la V^0 si mescolano; le combinazioni che si osservano sono il fotone (γ) e un vettore della forza debole chiamato Z^0. La simmetria $SU(2) \times U(1)$ si spezza spontaneamente. Questo evento ha l'effetto di conferire alle W^+, W^- e Z^0 grandi masse, lasciando invece il fotone privo di massa.

In eventi come questi si possono scoprire alcune relazioni sconcertanti tra la forza debole e l'elettromagnetismo. In primo luogo, le particelle W della forza debole possono portare la stessa quantità di carica debole e di carica elettrica. In secondo luogo, nella struttura dei singoletti e dei doppietti deboli si trova una curiosa relazione costante tra la carica debole e la carica elettrica. La carica elettrica di una particella è invariabilmente uguale alla somma della carica debole della particella e della carica elettrica media del singoletto o del doppietto al quale la particella appartiene. Chiamerò carica $U(1)$ tale carica media. Per i singoletti la carica $U(1)$ è semplicemente la carica elettrica della particella stessa e la regola è poco più che una tautologia, in quanto essa afferma che la carica elettrica è uguale alla carica elettrica, essendo sempre nulla la carica debole di una particella singoletto. Per i doppietti deboli, invece, la relazione tra le cariche è più interessante. È il caso di sottolineare che la relazione resta valida sia per doppietti di leptoni, dove le cariche mediate sono intere, sia per i doppietti dei quark, che sono formati da cariche frazionarie.

Come le altre cariche, la carica $U(1)$ è associata a una simmetria. Alla simmetria $U(1)$, estratta in questo modo dalle interazioni deboli, è associata una particella che chiamerò V^0. Come la W^0 e il fotone, la V^0 non ha né carica elettrica né carica debole. In effetti, dal momento che le cariche debole, $U(1)$ ed elettrica sono correlate, devono essere correlate anche le tre particelle portatrici. Risulta che la W^0 e la V^0 non si osservano in natura come stati puri, ma appaiono solo come miscele. Una di tali miscele della W^0 e della V^0 è il fotone; l'altra possibile combinazione è identificata nella particella chiamata Z^0. Sia il fotone sia la Z^0 mediano interazioni nelle quali una particella si trasforma in se stessa, in modo da lasciare inalterata la propria identità. Esse differiscono tuttavia per il fatto che il fotone accoppia solo

Nubi di particelle virtuali circondano un carica centrale puntiforme e alterano la sua risposta a una forza. Una carica elettrica positiva viene circondata da fotoni virtuali e da coppie virtuali elettrone-positone. I fotoni hanno un effetto scarso, ma le particelle cariche virtuali sono polarizzate, quindi le cariche negative virtuali si addensano attorno alla carica reale positiva, riducendone il valore efficace. Una carica di colore positiva è circondata da una nube di gluoni virtuali e di coppie virtuali quark-antiquark. I quark e gli antiquark si polarizzano come gli elettroni e i positoni, mentre i gluoni si comportano diversamente dai fotoni. Mentre il fotone è elettricamente neutro, alcuni gluoni hanno una carica di colore, quasi sempre della stessa polarità della carica reale. La carica di colore diffonde perciò nello spazio e la carica totale in un qualsiasi volume sferico diventa minore al diminuire del raggio della sfera. Il risultato finale di questi effetti è che l'interazione elettromagnetica diventa maggiore a piccola distanza, mentre l'interazione forte diventa minore. Anche l'interazione debole ha particelle portatrici di carica, e anch'essa diventa più debole a distanze piccole, anche se si indebolisce in minor misura di quanto fanno le interazioni forti.

particelle dotate di carica elettrica, mentre la Z^0 accoppia quelle con carica debole, tra cui i neutrini. Ricavando il fotone e le particelle *W* e *Z* dalla stessa teoria, la forza debole e l'elettromagnetismo vengono parzialmente unificati. La teoria risultante viene indicata con il prodotto dei gruppi che essa comprende: $SU(2)\times U(1)$.

Per analogia con la $U(1)$ dell'elettrodinamica quantistica e la $SU(3)$ della cromodinamica quantistica, ci si potrebbe aspettare che la $SU(2)$ e la $U(1)$ dell'interazione debole siano simmetrie esatte della teoria. In tal caso la carica debole dovrebbe conservarsi esattamente e le particelle *W* e *Z* dovrebbero essere prive di massa e con un raggio d'azione infinito. Invece, come abbiamo notato sopra, la carica debole non si conserva sempre; inoltre, si è osservato che la forza debole ha un raggio d'azione straordinariamente breve, di forse 10^{-15} centimetri. Il motivo è che la W^+, la W^- e la Z^0 hanno masse grandi, di quasi 100 GeV, pari a 100 volte la massa del protone. Che cosa accade della simmetria $SU(2)\times U(1)$ delle interazioni debole ed elettromagnetica in tali condizioni? Inoltre, se il fotone e le particelle *Z* e *W* sono strettamente correlati, come può il fotone restare privo di massa se le altre acquistano una grande massa?

La risposta a tali domande costituì un passo essenziale nella formulazione della teoria combinata delle forze debole ed elettromagnetica. La risposta oggi preferita è che la forza fondamentale sia davvero simmetrica e che in qualche ipotetico stato iniziale tutti i portatori della forza debole siano privi di massa. Ciò che non è simmetrico è il vuoto quantomeccanico, il mezzo in cui tale forza opera. La struttura del vuoto spezza spontaneamente la simmetria $SU(2)\times U(1)$, attribuendo una massa ai tre portatori della forza debole ma non al fotone. Una perdita di simmetria analoga si può osservare in un cristallo di sale. Gli ioni del cristallo non privilegiano alcuna direzione dello spazio rispetto a un'altra; gli ioni hanno simmetria rotazionale. Tuttavia, nella struttura reticolare del cristallo la simmetria viene spezzata e certe direzioni, quali quelle parallele alle linee reticolari, hanno uno stato particolare.

La stessa analogia può gettare luce su un'altra caratteristica di una simmetria spezzata spontaneamente. L'aspetto del mondo dipende dalla scala con la quale viene esaminato. Nel cristallo si possono distinguere tre scale di distanza. A distanze molto inferiori a 10^{-8} centimetri (le dimensioni di un atomo) si vede la struttura interna di un atomo, che risulta del tutto simmetrica e non influenzata dalla disposizione del cristallo. A distanze di circa 10^{-8} centimetri diventano importanti le forze responsabili della coesione degli atomi nel cristallo e i fenomeni osservati sono estremamente complessi. A distanze molto maggiori di 10^{-8} centimetri la geometria del cristallo appare in tutta la sua completezza e la simmetria rotazionale degli atomi viene visibilmente spezzata.

Una gerarchia di scale di distanza di questo genere si può definire in connessione con la rottura spontanea della simmetria $SU(2)\times U(1)$, ma la distanza critica è molto inferiore: circa 10^{-16} centimetri. A distanze molto inferiori a questa la simmetria si esprime pienamente. In un raggio d'azione così limitato le particelle pesanti *W* e *Z* vengono scambiate con la stessa rapidità dei fotoni privi di massa e in tal modo vengono efficacemente unificate la forza debole e quella elettromagnetica. Un altro modo di esprimere la stessa idea è quello di evidenziare una relazione tra distanza ed energia. Secondo il principio di indeterminazione, l'energia necessaria per sondare o risolvere una data distanza è inversamente proporzionale alla distanza. Un esperimento per esaminare la struttura di una particella a una distanza inferiore a 10^{-16} centimetri avrebbe dovuto essere eseguito a un'energia maggiore di 100 GeV. A questa energia le particelle *W* e *Z* possono essere create liberamente e la differenza di massa tra esse e il fotone è trascurabile; tutte le particelle intermediarie sono leggere in confronto all'energia dell'esperimento.

A una distanza di circa 10^{-16} centimetri cominciano a entrare in gioco i complessi fenomeni responsabili della rottura della simmetria $SU(2)\times U(1)$. Si osservano ancora le particelle *W* e *Z*, ma esse appaiono ben diverse dal fotone, perché le loro masse sono confrontabili all'energia dell'esperimento. A distanze ancora maggiori la simmetria tra il fotone e le particelle *W* e *Z* viene completamente a mancare; in realtà, l'energia è insufficiente per la creazione di una *W* o di una *Z* reali e quindi queste particelle non si possono osservare direttamente. Si possono rivelare solo i rari effetti a breve raggio d'azione del loro scambio virtuale (come il decadimento beta del neutrone).

Il concetto di rottura spontanea della simmetria risolve la questione della conservazione della carica debole. A un'energia molto maggiore di 100 GeV, dove si osserva direttamente la simmetria $SU(2)\times U(1)$, la massa di un quark o di un leptone è trascurabile; l'orientamento delle particelle è perciò praticamente costante e quindi la carica debole si conserva efficacemente. A bassa energia, dove la simmetria si spezza spontaneamente, la carica debole non si conserva, ma può scomparire nel vuoto quando una particella pesante cambia orientamento.

Tale teoria delle interazioni debole ed elettromagnetica è stata elaborata negli anni sessanta da Sheldon Lee Glashow e da Steven Weinberg della Harvard University e da Abdus Salam del Centro internazionale di fisica teorica di Trieste. Glashow fu il primo a ricavare la forma della teoria $SU(2)\times U(1)$, ma non sapeva come incorporare in essa le masse delle particelle *W* e *Z*. Weinberg e Salam scoprirono separatamente in seguito la forma della $SU(2)\times U(1)$ e applicarono il concetto di rottura spontanea della simmetria, formulando in tal modo una teoria coerente.

La teoria $SU(2)\times U(1)$ è solo un'unificazione parziale, perché comprende ancora due forze distinte, ciascuna con il proprio gruppo di simmetria e con la propria costante di accoppiamento. Il rapporto delle costanti di accoppiamento è un parametro libero, che deve essere scelto in modo da accordarsi con i dati sperimentali. Un'altra mancanza della teoria è che la carica elettrica è solo parzialmente quantizzata. La costruzione di doppietti di particelle correlate da trasformazioni $SU(2)$ richiede che tutte le differenze tra le cariche elettriche siano intere, in modo che ogni particella possa mutare la propria identità emettendo una W^+ o una W^-. La carica media dei doppietti, però, non è quantizzata. La carica media è la carica $U(1)$ definita in connessione con la carica elettromagnetica dell'elettrodinamica quantistica; come in tale teoria, non esiste alcun motivo fondamentale per confinare le cariche a valori interi. In realtà, nei doppietti formati da quark l'intervallo tra le cariche è un intero, ma le cariche effettive sono frazionarie.

L'impresa di costruire una teoria unificata delle forze debole, forte ed elettromagnetica non dovrebbe essere vista come un semplice tentativo di liberarsi del modello $SU(3)$ del colore o del modello $SU(2)\times U(1)$. Le teorie delle singole forze vanno troppo bene per essere smantellate. Ciò che una teoria unificatrice può fornire è una superstruttura nella quale possano essere inserite le teorie $SU(3)$ e

	CARICA ELETTRICA	CARICA DEBOLE	CARICA R — V	CARICA V — B
$d^{\text{ROSSO}}_{\text{DESTRA}}$	−1/3	0	+1/2	0
$d^{\text{VERDE}}_{\text{DESTRA}}$	−1/3	0	−1/2	+1/2
$d^{\text{BLU}}_{\text{DESTRA}}$	−1/3	0	0	−1/2
e^{+}_{DESTRA}	+1	+1/2	0	0
$\bar{\nu}_{\text{DESTRA}}$	0	−1/2	0	0
	= 0	= 0	= 0	= 0

La famiglia complessiva delle particelle elementari comprende sia leptoni sia quark. La famiglia ha cinque membri: tre quark destrorsi (i tre colori del quark *d*) e due antileptoni destrorsi (il positone e l'antineutrino). A ogni particella si assegnano valori della carica elettrica, della carica debole e di due cariche di colore: rosso meno verde (R – V) e verde meno blu (V – B). La terza carica di colore viene omessa perché ridondante. Per ogni tipo di carica la somma dei valori assegnati alle particelle è zero. Inoltre, nella teoria unificata le particelle con trasformazioni all'interno della famiglia trasportano ogni tipo di carica. Da questi due fatti deriva che tutte le cariche delle particelle sono quantizzate. La teoria unificata è associata al gruppo di simmetria *SU*(5).

	$d^{\text{ROSSO}}_{\text{DESTRA}}$	$d^{\text{VERDE}}_{\text{DESTRA}}$	$d^{\text{BLU}}_{\text{DESTRA}}$	e^{+}_{DESTRA}	$\bar{\nu}_{\text{DESTRA}}$
$d^{\text{ROSSO}}_{\text{DESTRA}}$	$G_1 + G_2 + \gamma + Z^0$	$G_{R\to V}$	$G_{R\to B}$	$X^{\text{ROSSO}}_{-4/3}$	$X^{\text{ROSSO}}_{-1/3}$
$d^{\text{VERDE}}_{\text{DESTRA}}$	$G_{V\to R}$	$G_1 + G_2 + \gamma + Z^0$	$G_{V\to B}$	$X^{\text{VERDE}}_{-4/3}$	$X^{\text{VERDE}}_{-1/3}$
$d^{\text{BLU}}_{\text{DESTRA}}$	$G_{B\to R}$	$G_{B\to V}$	$G_1 + G_2 + \gamma + Z^0$	$X^{\text{BLU}}_{-4/3}$	$X^{\text{BLU}}_{-1/3}$
e^{+}_{DESTRA}	$X^{\overline{\text{ROSSO}}}_{+4/3}$	$X^{\overline{\text{VERDE}}}_{+4/3}$	$X^{\overline{\text{BLU}}}_{+4/3}$	$\gamma + Z^0$	W^+
$\bar{\nu}_{\text{DESTRA}}$	$X^{\overline{\text{ROSSO}}}_{+1/3}$	$X^{\overline{\text{VERDE}}}_{+1/3}$	$X^{\overline{\text{BLU}}}_{+1/3}$	W^-	Z^0

La simmetria *SU*(5) comprende tutte le possibili transizioni tra le cinque particelle della famiglia complessiva. Le simmetrie delle singole forze sono incorporate dalla teoria *SU*(5) come sottogruppi: la simmetria *SU*(3) della forza forte è compresa nella matrice tre per tre in alto a sinistra, mentre la simmetria *SU*(2) della forza debole appare nella matrice due per due in basso a destra. La simmetria *U*(1) associata alla Z^0 appare negli accoppiamenti del fotone e della Z^0 lungo la diagonale. La teoria *SU*(5) postula 12 nuove particelle intermediarie, chiamate *X*, che mediano le transizioni tra quark e leptoni. La reciproca trasformabilità di questi è possibile solo in una teoria unificata.

TIPO DI CARICA

	ELETTRICA	DEBOLE	R — V	V — B
d^{ROSSO}_{DESTRA}	−⅓	0	+½	0
+ e^{+}_{DESTRA}	+1	+½	0	0
$u^{ROSSO}_{SINISTRA}$	+⅔	+½	+½	0
d^{VERDE}_{DESTRA}	−⅓	0	−½	+½
+ e^{+}_{DESTRA}	+1	+½	0	0
$u^{VERDE}_{SINISTRA}$	+⅔	+½	−½	+½
d^{BLU}_{DESTRA}	−⅓	0	0	−½
+ e^{+}_{DESTRA}	+1	+½	0	0
$u^{BLU}_{SINISTRA}$	+⅔	+½	0	−½
d^{ROSSO}_{DESTRA}	−⅓	0	+½	0
+ $\bar{\nu}_{DESTRA}$	0	−½	0	0
$d^{ROSSO}_{SINISTRA}$	−⅓	−½	+½	0
d^{VERDE}_{DESTRA}	−⅓	0	−½	+½
+ $\bar{\nu}_{DESTRA}$	0	−½	0	0
$d^{VERDE}_{SINISTRA}$	−⅓	−½	−½	+½
d^{BLU}_{DESTRA}	−⅓	0	0	−½
+ $\bar{\nu}_{DESTRA}$	0	−½	0	0
$d^{BLU}_{SINISTRA}$	−⅓	−½	0	−½
e^{+}_{DESTRA}	+1	+½	0	0
+ $\bar{\nu}_{DESTRA}$	0	−½	0	0
$e^{+}_{SINISTRA}$	+1	0	0	0
d^{VERDE}_{DESTRA}	−⅓	0	−½	+½
+ d^{BLU}_{DESTRA}	−⅓	0	0	−½
$\bar{u}^{\overline{ROSSO}}_{SINISTRA}$	−⅔	0	−½	0
d^{ROSSO}_{DESTRA}	−⅓	0	+½	0
+ d^{BLU}_{DESTRA}	−⅓	0	0	−½
$\bar{u}^{\overline{VERDE}}_{SINISTRA}$	−⅔	0	+½	−½
d^{ROSSO}_{DESTRA}	−⅓	0	+½	0
+ d^{VERDE}_{DESTRA}	−⅓	0	−½	+½
$\bar{u}^{\overline{BLU}}_{SINISTRA}$	−⅔	0	n	+½

La famiglia di 10 particelle è costruita formando tutte le possibili coppie dei cinque stati che costituiscono la più semplice famiglia della *SU*(5). Tutti i cinque membri della famiglia sono destrorsi, ma, se sono combinati a coppie, danno origine a stati sinistrorsi. Come nella costruzione di antiquark da coppie di quark, il processo non va inteso come un raggruppamento fisico; una particella sinistrorsa non è in realtà uno stato legato di due particelle destrorse. Ciononostante, il procedimento di costruzione fornisce tutte le cariche degli stati sinistrorsi. Con altre due famiglie delle stesse dimensioni e di struttura simile, tutte le particelle della prima generazione trovano posto nella teoria e non vi sono posti vuoti.

$SU(2) \times U(1)$. La superstruttura assumerebbe la forma di una più vasta simmetria in cui appaiano strettamente correlati i quark e i leptoni.

La ricerca di una tale simmetria più estesa deve cominciare con la ricerca di un gruppo più ampio, che comprenda come strutture componenti sia la $SU(3)$ sia la $SU(2) \times U(1)$. Molti gruppi hanno tale proprietà, ma uno di essi presenta numerosi vantaggi che lo evidenziano tra tutti gli altri: si tratta del gruppo $SU(5)$, il gruppo di tutte le possibili trasformazioni di cinque oggetti distinti, o di una matrice cinque per cinque. È il più piccolo gruppo semplice che possa sistemare le simmetrie $SU(3)$ e $SU(2) \times U(1)$, e penso che possa essere il più completo fra i gruppi di simmetria della natura. Una teoria unificata basata sulla simmetria $SU(5)$ è stata elaborata nel 1973 da Glashow e da me.

Nella rappresentazione più semplice del gruppo $SU(5)$ i cinque oggetti sono le componenti destrorse del quark *d* in ciascuno dei colori rosso, verde e blu, la componente destrorsa del positone e la componente destrorsa dell'antineutrino elettronico (che possiede soltanto una componente destrorsa). A ognuna delle cinque particelle è assegnato un valore per ciascuna delle quattro cariche indipendenti: la carica elettrica, la carica debole e due cariche di colore, che supporrò siano R − V e V − B.

Ventiquattro particelle intermedie forniscono tutte le possibili transizioni tra questi cinque stati della materia. Quattro delle particelle sono il fotone, la Z^0 e i gluoni G_1 e G_2, che sono direttamente associate alle quattro cariche fondamentali. Tali particelle, non avendo carica, possono partecipare solo a quelle interazioni nelle quali l'identità di una particella non cambia. Delle restanti 20 particelle intermediarie, otto ci sono già familiari. Si tratta della W^+ e della W^-, che possono trasformare un positone in un antineutrino e viceversa, e dei sei gluoni che trasformano i colori dei quark. Con l'aggiunta di queste 12 particelle portatrici si possono spiegare tutte le interazioni osservate finora in natura. Il gruppo $SU(5)$ comprende però altre 12 particelle necessarie per conferire alla teoria la massima simmetria possibile. Le 12 particelle supplementari sono indicate con *X* e la loro funzione è quella di mediare la reciproca conversione di leptoni e quark. Ogni particella *X* trasporta carica di colore, carica debole e carica elettrica; le cariche elettriche hanno valori più o meno 1/3 e più o meno 4/3.

Come nella distribuzione delle cariche di colore in $SU(3)$, la tabella delle assegnazioni di cariche nella teoria $SU(5)$ presenta qualche sorprendente regolarità. Per ogni tipo di carica, la somma delle cariche assegnate alle cinque particelle è zero. Per esempio, ognuno dei tre colori dei quark ha una carica elettrica − 1/3, ma queste cariche vengono bilanciate dalla carica elettrica + 1 del positone. Un'osservazione correlata è che tutte le quattro varietà di carica vengono trasportate da almeno qualcuna delle particelle intermediarie della $SU(5)$. I gluoni hanno colore, la W^+ e la W^- hanno sia carica debole sia carica elettrica e le particelle *X* trasportano tutti e quattro i tipi di carica.

Da questi fatti si può dedurre che tutte le cariche sono necessariamente quantizzate. Tutte le cariche elettriche devono essere multiple di 1/3; se si accettasse nella famiglia qualche particella con carica diversa, le particelle portatrici della $SU(5)$ non potrebbero essere né emesse né assorbite da essa senza violare il principio di conservazione della carica. Inoltre, non è solo il minimo intervallo tra le cariche a essere fisso; gli effettivi valori delle cariche vengono determinati imponendo che la carica totale sia nulla. Questa è in fondo una spiegazione della quantizzazione della carica elettrica. Lo stesso requisito spiega la esatta commensurabilità delle cariche dei leptoni e dei quark, che implica a sua volta l'esatta neutralità dell'atomo. Per concludere, la sorprendente coincidenza che tutti i sistemi di particelle a colore neutro hanno carica elettrica intera deriva dalla organizzazione della famiglia.

Che cosa c'è da dire sulle restanti particelle della prima generazione? Uno degli aspetti più eleganti della teoria $SU(5)$ è che le cinque particelle destrorse della famiglia più piccola della $SU(5)$ si possono combinare a coppie per produrre una famiglia di 10 particelle sinitrorse. Questi 10 stati costituiscono la più semplice rappresentazione del gruppo. Si tratta delle componenti sinistrorse del quark *d*, del quark *u* e dell'antiquark $\bar{u}$ (ciascuno in tre colori) e del positone. Come nel «montaggio» di antiquark e gluoni a partire dal tripletto fondamentale di colori dei quark, questo processo non dovrebbe essere interpretato come una ricetta fisica per costruire le particelle. Un leptone o un quark sinistrorso non è effettivamente composto da due particelle destrorse in uno stato legato. Cionondimeno, i 10 modi possibili di formare coppie dei cinque stati destrorsi forniscono tutte le cariche corrette per le particelle sinistrorse.

Questo schema di composizione a coppie fornisce correttamente anche le transizioni possibili per ogni particella. E ciò che è ugualmente importante è che le transizioni non osservate non sono permesse dalla struttura del gruppo. Ogni famiglia di particelle è chiusa: ogni transizione permessa dà origine a un'altra particella della stessa famiglia e non è possibile alcun'altra transizione.

La famiglia di cinque stati destrorsi e quella di 10 stati sinistrorsi contengono complessivamente 15 particelle. Si possono costruire altre due famiglie di forma leggermente differente per sistemare le restanti 10 particelle destrorse e cinque sinistrorse, che sono le antiparticelle dei 15 stati delle prime due famiglie. Quindi tutti i 30 stati elementari della prima generazione hanno il loro posto nella teoria e non esistono posti vuoti. Rappresentazioni equivalenti delle generazioni di grado più elevato si possono costruire sostituendo il muone o il leptone tau all'elettrone, il quark *s* o *b* al quark *d*, e così via.

Mi si consenta di rivedere ciò che è stato fatto finora. Dapprima è stato scelto il gruppo $SU(5)$ come il più piccolo gruppo capace di comprendere sia SU(3) sia $SU(2) \times U(1)$. Successivamente sono state selezionate cinque componenti destrorse di particelle come membri della famiglia $SU(5)$ più semplice. Le restanti componenti dovevano allora trovar posto in qualche altra famiglia della $SU(5)$, come è avvenuto. Senza nulla trascurare, esse hanno trovato posto nella famiglia immediatamente più semplice. Inoltre, la composizione della famiglia derivata è stata specificata con un semplice procedimento per accoppiare le particelle. È importante sottolineare che questo procedimento di accoppiamento non doveva necessariamente funzionare; in molti gruppi diversi dalla $SU(5)$ non avrebbe funzionato. Questo rappresenta il primo, il più semplice e per alcuni aspetti il più notevole successo della teoria $SU(5)$.

Il significato più evidente dell'unificazione $SU(5)$ è che i leptoni e i quark non sono più inconciliabilmente diversi. Essi sono invece membri di una sola famiglia e un quark può essere trasformato in un leptone (o viceversa) come pure un quark può essere trasformato in un altro quark o un leptone in un altro leptone. Un'ulteriore conseguenza dell'unificazione è che le forze debole, forte ed elettromagnetica dovrebbero avere tutte la stessa intensità, o la stessa costante di accoppiamento. Nessuna di tali previsioni è soddisfatta nel mondo delle attuali osservazioni. Tra i milioni di interazioni di particelle elementari registrate dai fisici non è stato osservato un solo caso di trasformazione quark-leptone. Inoltre, le costanti di accoppiamento delle tre forze differiscono per fattori grandi: la forza forte è circa 100 volte più intensa dell'elettromagnetismo e la forza debole si trova a metà strada tra le due. Se quindi $SU(5)$ è una simmetria della natura, si tratta evidentemente di una simmetria malamente spezzata.

La simmetria potrebbe essere spezzata da un meccanismo simile a quello che spezza la simmetria $SU(2)\times U(1)$ delle forze debole ed elettromagnetica. In questo modo alle particelle *X* si dovrebbe dare una grande massa e gli effetti dello scambio di una particella *X* dovrebbero essere totalmente eliminati. Invece, nella $SU(5)$ la rottura dovrebbe comparire a un'energia molto maggiore o in modo equivalente, a una distanza molto minore che nella $SU(2)\times U(1)$. Tale distanza è la scala di unificazione, la distanza a cui diventa manifesta la completa simmetria della teoria.

Con la teoria $SU(5)$ si può ipotizzare come dovrebbe apparire il mondo nelle varie scale di distanze o di energia. In un esperimento che abbia sondato distanze molto minori della scala di unificazione, l'invarianza di gauge $SU(5)$ del mondo dovrebbe immediatamente manifestarsi. Tutte le interazioni, comprese le trasformazioni quark-leptone e quark-antiquark, dovrebbero avere lo stesso peso, dal momento che tutte le particelle intermediarie della $SU(5)$ (il fotone, i gluoni, le particelle *W* e *Z* e le particelle *X*) dovrebbero venir create all'incirca con la stessa probabilità. Le masse delle particelle *W*, *Z* e *X* le farebbero a malapena distinguere dal fotone e dai gluoni, essendo piccole in confronto all'energia dell'esperimento.

A distanze prossime alla scala di unificazione si dovrebbe osservare la fisica complessa associata alla rottura spontanea della simmetria $SU(5)$. Dovrebbero venire emesse particelle *X*, ma la loro massa le renderebbe molto diverse da tutte le altre particelle. A distanze molto maggiori della scala di unificazione (ma sempre molto minori di 10^{-16} centimetri) la simmetria $SU(5)$ dovrebbe essere quasi completamente nascosta. Le particelle *X* non potrebbero più essere create come particelle reali, e quindi i leptoni e i quark dovrebbero venire segregati in famiglie separate, con poche possibilità di comunicazione. D'altra parte, la simmetria $SU(2)\times U(1)$ resterebbe inalterata e si potrebbero fare poche distinzioni tra le interazioni debole ed elettromagnetica. A distanze maggiori di 10^{-16} centimetri dovrebbe spezzarsi anche la simmetria $SU(2)\times U(1)$ e vi sarebbero tre forze distinte.

La rottura spontanea della simmetria può anche spiegare le disparità tra le costanti di accoppiamento. L'elemento cruciale nella spiegazione è l'effetto delle particelle virtuali nel vuoto che circonda una carica puntiforme. Ricordiamo che nel caso della forza forte la nube di gluoni virtuali che avviluppa un quark si estende efficacemente al di fuori della carica di colore, con il risultato che la costante di accoppiamento diminuisce quando viene misurata a una distanza minore. Le particelle virtuali *W* hanno un effetto del genere su una carica debole, anche se l'effetto è alquanto minore in quanto le particelle *W* che trasportano la carica debole sono in numero inferiore a quello dei gluoni che trasportano la carica di colore. Del resto, nella teoria $U(1)$, il fatto che la Z^0 sia priva di carica origina un fenomeno del tutto differente. A causa della polarizzazione di elettroni e di positoni virtuali la costante di accoppiamento della $U(1)$ aumenta alle brevi distanze.

Da questi comportamenti delle costanti di accoppiamento si può trarre una sempli-

ce conclusione. A grandi distanze la costante di accoppiamento $SU(3)$ della forza forte è massima, ma diminuisce anche con la massima rapidità, mentre la costante $SU(2)$ della forza debole è minore e diminuisce più lentamente; la costante $U(1)$ è la minima, ma aumenta al diminuire della distanza. Sembra quindi che possa esistere una distanza alla quale tutte e tre le costanti hanno all'incirca lo stesso valore.

A distanze inferiori a 10^{-16} centimetri tutte le costanti di accoppiamento sono decisamente piccole e si può calcolare il modo in cui esse variano con la distanza o con l'energia. È possibile calcolarne i valori a distanze sempre più piccole; la distanza alla quale le tre costanti convergono è la scala di unificazione. Se si seguono in questo modo due delle tre costanti, si può prevedere a qualsiasi energia il valore della terza costante. Tali calcoli sono stati eseguiti utilizzando come dato la costante di accoppiamento forte e la costante di accoppiamento elettromagnetica; quest'ultima risulta ancora una combinazione delle costanti di accoppiamento $U(1)$ ed $SU(2)$. I risultati forniscono valori della scala di unificazione e del rapporto tra le costanti $U(1)$ e $SU(2)$, un parametro arbitrario delle teorie non unificate.

Nel 1974 ho eseguito un tale calcolo con Helen R. Quinn (che lavora oggi allo Stanford Linear Accelerator Center) e Weinberg per una classe di teorie unificate che comprende la $SU(5)$. Abbiamo ottenuto per la scala di unificazione una distanza di circa 10^{-29} centimetri e un valore di 0,2 per il rapporto tra $U(1)$ e $SU(1)$. In quell'epoca i risultati non erano incoraggianti, perché le misure del rapporto suggerivano un valore di circa 0,35. Dopo di allora, misure perfezionate del rapporto delle costanti di accoppiamento hanno dato valori inferiori. Attualmente si ritiene che il rapporto sia uguale a 0,22 più o meno 0,02, in accordo con il risultato teorico.

Un'altra previsione del modello $SU(5)$ che può essere verificata a energie accessibili è stata elaborata nel 1977 da Andrzej Buras, John Ellis, Mary K. Gaillard e Demetres V. Nanopoulos dell'Organizzazione europea per le ricerche nucleari (CERN) di Ginevra. Essi hanno trovato che nella versione più semplice della teoria $SU(5)$ si può calcolare il rapporto tra la massa del quark *b* e quella del leptone tau. Come nel caso delle costanti di accoppiamento, si prevede che le masse siano uguali in corrispondenza alla scala di unificazione, ma a distanze maggiori il quark *b* è più pesante a causa della sua carica di colore. Si è calcolato che il rapporto tra le masse a bassa energia è circa di 3:1. La massa della tau è nota: essa è pari a poco meno del doppio della massa del protone. La massa del quark *b* non è altrettanto sicura perché il quark non può essere esaminato da solo. La migliore stima attuale è di circa cinque volte la massa del protone, il che porta il rapporto delle masse a 5:2.

La scala di unificazione di 10^{-29} centimetri è una distanza straordinariamente piccola. (Se un singolo protone venisse gonfiato fino alle dimensioni del Sole, la scala di unificazione sarebbe ancora inferiore a un micrometro.) Nell'unificazione $SU(5)$ è implicata l'ipotesi che non si incontrino nuovi principi fisici nell'intero intervallo di distanze tra 10^{-16} e 10^{-29} centimetri; in particolare, si deve supporre che resti invariato il modo in cui variano con la distanza le costanti di accoppiamento. Una tale ipotesi è ovviamente scomoda, ma non è del tutto implausibile. Esiste già una piccola scala di distanze alla quale sono previsti nuovi fenomeni. A circa 10^{-33} centimetri la gravitazione può acquistare la stessa intensità delle altre forze e quindi qualsiasi teoria che descriva eventi su questa scala deve comprendere la gravitazione. Mi è parso incoraggiante il fatto che la scala di unificazione, pur essendo piccola, sia ancora 10 000 volte maggiore di 10^{-33} centimetri.

Una distanza di 10^{-29} centimetri corrisponde a un'energia di circa 10^{15} GeV, pari a circa 10^{15} volte la massa del protone. Le particelle *X* devono avere una massa equivalente a questa energia. Per amore di confronto, le più pesanti particelle che si possono oggi creare con gli acceleratori di particelle hanno una massa di circa 10 GeV. Con una nuova generazione di acceleratori, attualmente in fase di studio, si spera di raggiungere i circa 100 GeV necessari per la creazione di particelle *W* e *Z*. Per poter creare particelle *X* l'energia dovrebbe venire ulteriormente aumentata di 13 ordini di grandezza, il che pare del tutto improbabile da realizzare.

Pur non essendo mai possibile mostrare una particella *X* reale in laboratorio, l'esistenza di tali particelle si potrebbe dimostrare rivelando eventi nei quali viene scambiata una particella *X* virtuale. Tali scambi a loro volta dovrebbero essere estremamente rari, dal momento che potrebbero avvenire soltanto quando a due particelle elementari capita di passare a meno di 10^{-29} centimetri una dall'altra. Tuttavia, anche nel campo degli eventi più comuni, lo scambio di una particella *X* dovrebbe essere immediatamente identificato, perché le particelle *X* possono fare qualcosa che nessuna delle altre particelle può fare: possono trasformare un quark in un leptone o un quark in un antiquark. Questo processo mette in discussione la stessa stabilità della materia.

Si prevede che l'unificazione delle forze nella teoria $SU(5)$ diventi evidente a energia estremamente alta o, in modo equivalente, a distanza piccolissima. L'entità intrinseca di ciascuna forza è misurata da una costante di accoppiamento adimensionale associata alla simmetria fondamentale. A causa degli effetti delle particelle virtuali che circondano una carica, l'interazione forte $SU(3)$ diminuisce di intensità al diminuire della distanza tra le particelle. Anche l'interazione $SU(2)$ diventa più debole a piccola distanza, ma con minor rapidità. Invece, l'interazione $U(1)$ diventa più intensa. L'estrapolazione dei valori delle costanti, misurati a distanza relativamente grande, fa ritenere che le curve convergano a una distanza di circa 10^{-29} centimetri, equivalente a un'energia di 10^{15} gigaelettronvolt. Per tale distanza ed energia tutte le forze dovrebbero avere la stessa intensità e tutte le transizioni tra le particelle elementari dovrebbero essere equiprobabili. A distanze maggiori di 10^{-16} centimetri la simmetria $SU(2) \times U(1)$ si spezza e le forze $SU(2)$ e $U(1)$ cessano di esistere come entità separate. La simmetria $U(1)$ dell'elettromagnetismo sopravvive, ma la forza debole appare solamente attraverso lo scambio virtuale di particelle *W* e *Z*.

Le interazioni mediate da particelle *X* differiscono da tutte le altre interazioni, nel senso che violano la conservazione di una quantità chiamata numero barionico. Il numero barionico di una particella qualsiasi può essere definito come un terzo del numero di quark meno un terzo del numero di antiquark. Quindi il protone e tutte le altre particelle formate da tre quark hanno un numero barionico + 1, mentre il mesone pi greco e altre particelle

Il decadimento del protone è l'inevitabile conseguenza delle transizioni $SU(5)$ che trasformano un quark in un leptone o un quark in un antiquark. La natura di tale trasformazione è illustrata nel diagramma a sinistra. Dapprima un quark rosso destrorso *d* emette una particella *X* rossa con carica elettrica – 4/3; il quark si trasforma quindi in un positone destrorso. Successivamente la particella *X* viene assorbita da un quark verde sinistrorso *u* che diventa perciò un antiquark sinistrorso antiblu $\bar{u}$. Il diagramma di destra illustra il risultato del processo quando esso avviene in un protone che forma il nucleo di un atomo di idrogeno.

Il protone ha una composizione in quark *uud*. In seguito allo scambio della particella *X*, il quark *d* diventa un positone, che alla fine entra in collisione con l'elettrone dell'atomo (o con qualche altro elettrone), ed entrambe le particelle si annichilano con un fiotto di fotoni di alta energia. Ciò che rimane del protone dopo lo scambio della particella *X* sono un quark *u* e un antiquark $\bar{u}$, che costituiscono insieme un mesone pi neutro. Anche il mesone pi neutro decade in fotoni di alta energia. Il risultato finale dello scambio è la trasformazione dell'intera massa dell'atomo di idrogeno in radiazione elettromagnetica.

formate da un quark e da un antiquark hanno numero barionico 0. Naturalmente anche i leptoni hanno un numero barionico 0, dal momento che non comprendono né quark né antiquark. Nelle interazioni forte, debole ed elettromagnetica il numero barionico totale non può mai cambiare. Se la conservazione del numero barionico fosse una legge assoluta della natura, il protone non potrebbe mai decadere, perché è la particella più leggera con numero barionico non nullo. L'unificazione $SU(5)$ prevede proprio tale decadimento.

Una possibile violazione della conservazione del numero barionico è quella di un protone che forma il nucleo di un atomo di idrogeno. Il protone è formato da quark, *u*, *u* e *d*, con un quark per colore. Se due dei quark vengono a trovarsi a una distanza inferiore a 10^{-29} centimetri, una particella *X* può passare in mezzo a loro. Per esempio, un quark rosso destrorso può emettere una *X* con carica elettrica –4/3 e cariche di colore corrispondenti al colore rosso. Il quark *d*, avendo perso la propria carica di colore e avendo cambiato la propria carica elettrica da – 1/3 a + 1, diventerebbe perciò un positone. La particella *X* potrebbe essere assorbita da un quark verde sinistrorso *u* che si trasformerebbe in un antiquark sinistrorso $\bar{u}$ di colore antiblu. Il nuovo antiquark $\bar{u}$ potrebbe combinarsi con il quark *u* restante per formare un mesone pi neutro. I numeri barionici sia del positone sia del mesone pi sono nulli, quindi il numero barionico totale varia da + 1 a 0.

Se tale evento fosse osservato in laboratorio, lo scambio della particella *X* non potrebbe essere direttamente rivelato. Tutto ciò che si potrebbe osservare sarebbe il decadimento di un protone in un positone e in un mesone pi neutro. Inoltre, la sequenza di eventi non finirebbe qui. Il positone incontrerebbe in seguito un elettrone (forse l'elettrone dell'atomo di idrogeno) ed essi si annichilerebbero a vicenda, dando origine a raggi gamma, o fotoni di alta energia. Il quark *u* e l'antiquark $\bar{u}$ del mesone pi neutro finirebbero con l'annichilarsi a vicenda allo stesso modo, liberando più raggi gamma. Il risultato finale è che un atomo di idrogeno decade in uno stato di radiazione pura. Questo processo rappresenta una trasformazione di materia in energia molto più efficiente di quello della fissione nucleare o della fusione termonucleare. La fusione di atomi di idrogeno per formare elio libera meno dell'1 per cento della loro massa sotto forma di energia, mentre questo processo libera il 100 per cento della massa.

La brusca scomparsa di un protone, e quindi di un atomo, è un evento che probabilmente accade molto di rado, altrimenti sarebbe già stato osservato da molto tempo. In realtà, ci si deve aspettare una bassa frequenza perché raramente le particelle entrano all'interno della distanza alla quale è possibile lo scambio di una particella *X*. Quinn, Weinberg e io abbiamo utilizzato il nostro calcolo della scala di unificazione per stimare la velocità di decadimento del protone. La nostra stima è stata da allora perfezionata da molti altri, tra i quali Buras, Ellis, Gaillard e Nanopoulos del CERN, Terrence J. Goldman e Douglas A. Ross del Cal Tech e William J. Marciano della Rockefeller University. La stima attuale è che la vita media del protone sia di circa 10^{31} anni.

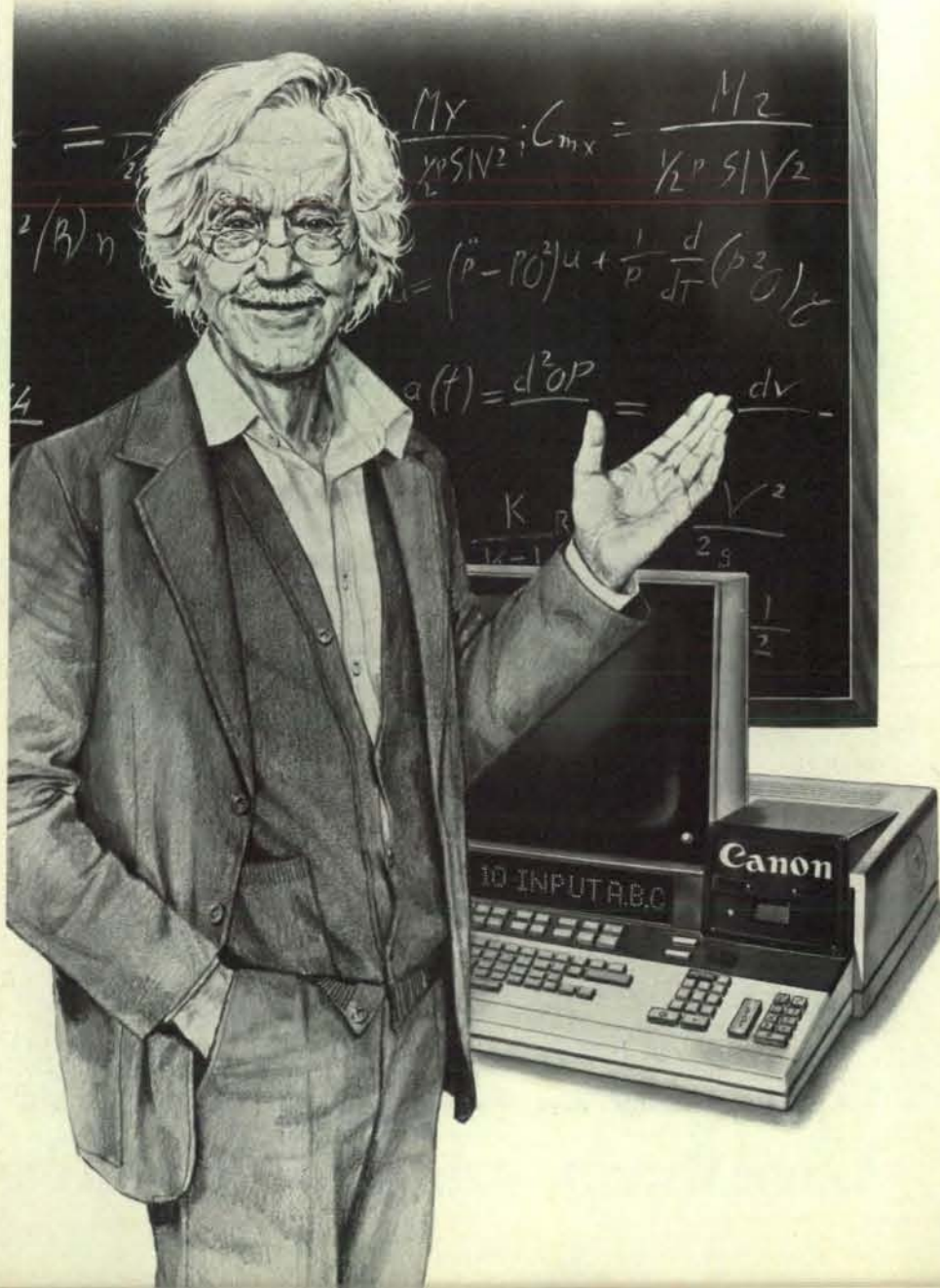
Per verificare l'unificazione $SU(5)$ è ovviamente scomodo aspettare 10^{31} anni perché un dato protone decada. L'età dell'universo dopo il big bang è di soli circa 10^{10} anni. Tuttavia la ricerca del decadimento del protone non è priva di speranza. Una vita media di 10^{31} anni implica che in un insieme di 10^{31} protoni si dovrebbe poter osservare un decadimento all'anno. In 1000 tonnellate di materia vi sono circa 5×10^{32} tra protoni e neutroni, quindi ci si può aspettare che circa 50 di essi decadano ogni anno. Quindi la strategia per la rivelazione di eventi che violano la conservazione del numero barionico è quella di tener d'occhio tutto ciò che accade in almeno 1000 tonnellate di materia per parecchi anni e di discriminare i protoni e i neutroni che decadono da eventi più comuni.

Numerosi gruppi di ricercatori stanno progettando esperimenti su questa scala. Gli esperimenti saranno eseguiti in profondità nel terreno o sott'acqua per ridurre al minimo il numero dei raggi cosmici che attraversano il campione di materia. I raggi cosmici possono dare origine a interazioni che potrebbero essere scambiate per decadimenti del protone. Un esperimento deve essere allestito in una miniera

di sale nei pressi di Cleveland, un altro in una miniera di argento nello Utah e un terzo in una miniera di ferro nel Minnesota. Esperimenti in scala leggermente inferiore sono in fase di studio per due tunnel sotto le Alpi, mentre esperimenti su scala ancora minore sono già in corso in miniere d'oro nel South Dakota e in India.

L'energia necessaria per creare particelle X reali può essere decisamente al di là delle capacità delle macchine create dall'uomo e in realtà non ci può essere oggi alcun processo, in nessun punto dell'universo, che sia in grado di generare un'energia così alta. Può darsi, però, che in un'epoca precedente le particelle X siano state comuni. Circa 10^{-40} secondi dopo il big bang le dimensioni dell'universo erano confrontabili con la scala di unificazione. L'universo allora era così caldo (circa 10^{18} kelvin) che tutte le particelle avevano energie confrontabili con la massa della X. Di conseguenza la simmetria $SU(5)$ stava appena cominciando a spezzarsi e le trasformazioni quark-leptone erano frequenti come qualsiasi altra interazione. Non si poteva fare alcuna distinzione fondamentale tra quark e leptoni o tra forze forte, debole ed elettromagnetica: c'era soltanto un tipo di materia e una sola forza.

Resti di quell'epoca di evidente simmetria possono essere ancora presenti nell'universo del giorno d'oggi; in un certo senso l'universo odierno costituisce il residuo. Un vecchio enigma dell'astrofisica si chiede perché l'universo è fatto di materia anziché di antimateria. In effetti, forse la più semplice previsione è che ci siano quantità uguali di materia e di antimateria, che alla fine si annichilano a vicenda per lasciare un universo formato da sola radiazione. L'unificazione $SU(5)$ presenta un'ipotetica spiegazione dell'apparente predominio della materia. È possibile che il libero interscambio di particelle X in un breve periodo successivo alla rottura della simmetria $SU(5)$ abbia creato più quark che antiquark e quindi più barioni che antibarioni.

L'affascinante ipotesi che i processi che violano la conservazione del numero barionico possano essere responsabili dell'eccesso di barioni fu avanzata per la prima volta dal fisico sovietico Andrei Sakharov nel 1967. Più recentemente Motohiko Yoshimura della Università Tohoku ha suggerito che la violazione del numero barionico prevista dalle teorie unificate ha tutte le proprietà necessarie per spiegare l'eccesso osservato. L'idea, dopo di allora, è stata rielaborata da molti altri, tra i quali Ellis, Gaillard, Nanopoulos, Weinberg, Savas Dimopoulos e Leonard Susskind della Stanford University e da Sam B. Treiman, Anthony Zee e Wilczek di Princeton. Essi hanno dimostrato che un eccesso di barioni può nascere soltanto se i processi che violano la conservazione del numero barionico appaiono differenti quando vengono visti a ritroso nel tempo. Tale condizione è contemplata nella teoria $SU(5)$. Così, una prova a favore dell'unificazione $SU(5)$, pur costituendo un'evidenza indiretta e circostanziata, è la stessa esistenza della materia.

Artemia, un nuovo animale d'allevamento

La maricoltura e il commercio di questo crostaceo potranno trarre notevoli vantaggi dalla ricerca genetica diretta alla selezione di ceppi dotati di opportuni caratteri economici

di Claudio Barigozzi

Anche se già Linneo si era accorto dell'esistenza di questo piccolo crostaceo che predilige le acque salate (addirittura verso la saturazione, nei deserti bacini delle saline dove quasi nessun altro vivente riesce a resistere) e lo aveva chiamato *Cancer salinus* (poi ribattezzato *Artemia* da Gerald Leach), per molti decenni esso pareva destinato a essere puro oggetto di ricerca teorica. Nonostante le caratteristiche dell'artemia avessero attirato l'attenzione di citologi, zoologi ed embriologi, essa non era mai diventata un oggetto di ricerche di moda come la drosofila. Pochi studiosi isolati la prediligevano, come Cesare Artom che dedicò quasi tutta la sua attività a questo crostaceo, dal 1906 al 1934 quando morì.

Eppure quasi tutti i problemi salienti della biologia animale, con al centro la citogenetica, hanno nell'artemia aspetti singolari e interessanti. Fu così che biologi di vari paesi lavorarono sull'artemia, spesso raccogliendola dalle saline, raramente, come fu realizzato da me fin dal 1939 e da S. T. Bowen a San Francisco alcuni anni dopo, allevandola in laboratorio. Non so chi sia stato il primo a pensare che le cosiddette uova durevoli o cisti, che si trovavano a miriadi nella terra secca delle saline dopo la raccolta del sale, potessero essere utilizzate per nutrire i pesci ornamentali da appartamento: certo fin dal 1960 era in atto una raccolta estesa, che portava a mettere in vendita tubetti pieni di cisti, che, messe in acqua salata, schiudono: ne fuoriesce una larvetta rosa, il cosiddetto *nauplius*, che i pesciolini mangiano assai volentieri.

Più tardi, fra il 1970 e il 1980, con lo sviluppo dell'acquacoltura e in particolare della maricoltura, cioè sotto la spinta di rifornire il mercato di crostacei e pesci d'allevamento, l'artemia venne presa in considerazione come nutrimento ideale. Ora l'artemia viene allevata in massa, sollevando tutti quei problemi di fisiologia e di genetica che comportano tutte quelle forme viventi che l'uomo tende a trasformare e a migliorare secondo le proprie finalità pratiche. Si può dire che quello dell'artemia è il più recente esempio di allevamento scoperto dall'uomo e come tale va considerato.

Vedremo tra poco che il genere *Artemia* comprende diverse specie, solo alcune delle quali sono ben diagnosticate. Da quando il genere fu denominato *Artemia* fino al 1968 si ritenne che tutti gli individui studiati appartenessero a una specie sola, *A. salina*, anche se furono descritte numerose forme locali e varianti morfologiche, specialmente in rapporto alla salsedine. Tutte queste varianti possono, tuttavia, essere dimenticate, sia perché scientificamente di dubbio significato, sia perché superate dalle recenti ricerche sulla speciazione.

L'artemia è un piccolo organismo lungo 2-3 centimetri, che, osservato anche a piccolo ingrandimento, risulta formato da tre parti. La prima (il capo) porta due paia di antenne (le seconde sono molto grandi nel maschio) e tre occhi, di cui uno mediano. La seconda è segmentata e porta 11 paia di arti forniti di appendici lamellari che servono alla respirazione. La terza è filiforme e consta essenzialmente dell'intestino che è rettilineo: all'inizio di questa ultima parte si trova nella femmina l'ovisacco o utero, nel quale si raccolgono le cisti, che sono sferoidali e rivestite da un guscio bruno. La larva (nauplius) consta di due sole parti: il capo con l'unico occhio mediano, le due paia di antenne e un paio di appendici che diventeranno mandibole e, posteriormente, una porzione senza appendici e non segmentata. Per merito di A. Brauer nel 1896 e di C. Artom nel 1906, l'artemia divenne il primo esempio di specie animale nella quale esistono due numeri cromosomici multipli, uno (42) caratteristico delle forme bisessuate e l'altro (84) delle forme partenogenetiche. Queste ultime furono trovate nelle saline di Capodistria mentre le prime sono tipiche delle saline di Cagliari.

Rappresentazione schematica di artemia femmina durante due cicli di produzione di uova. Il tubo intestinale è indicato solo nella prima figura, nelle successive i punti neri indicano le uova, dapprima nei due ovari, poi nelle due tasche dell'ovisacco e quindi raccolte in una massa mediana prima dell'espulsione. Da questo momento riprende il ciclo successivo.

Il fenomeno delle serie cromosomiche multiple (poliploidia) fu scoperto negli animali proprio nell'artemia. La partenogenesi è da considerarsi indefinita, nel senso che, a differenza di altri casi noti negli insetti, ogni individuo femmina produce uova non fecondate che daranno origine solo a individui femminili. In tempi recenti si sono osservati anche maschi nati da femmine partenogenetiche; si tratta però di un fenomeno non del tutto spiegato, raro e comunque discontinuo.

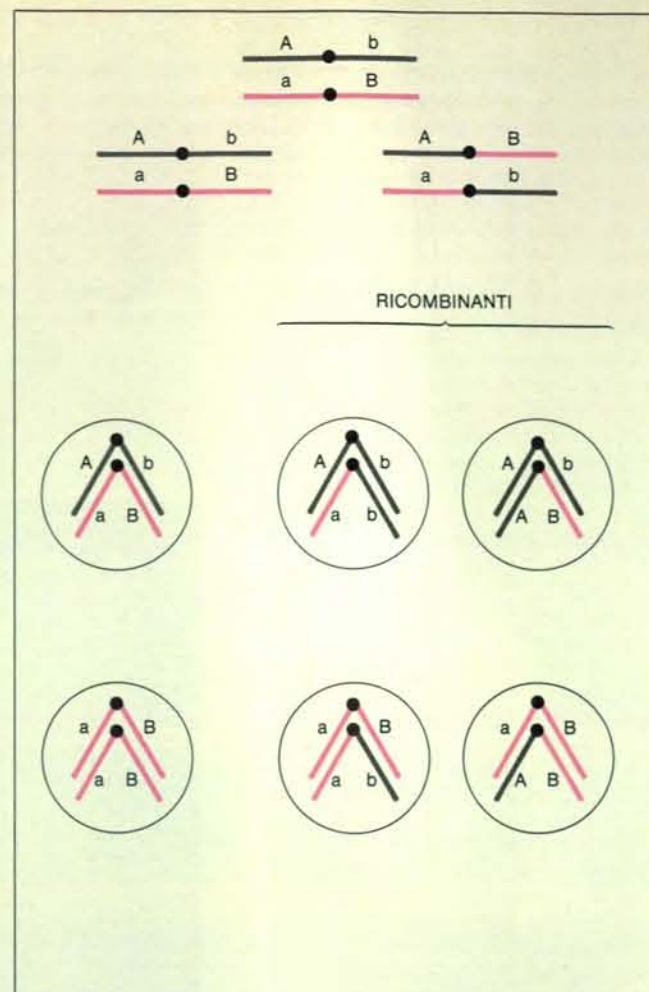
Confrontiamo ora la situazione genetica nelle artemie bisessuate e nelle partenogenetiche, alla luce specialmente dei fenomeni cromosomici. Nelle prime, i 42 cromosomi (fra i quali secondo R. Stefani si possono distinguere i cromosomi sessuali, XX nel maschio e XY nella femmina, fatto geneticamente perfettamente convalidato) vanno incontro alle normali meiosi maschile e femminile (che portano alla formazione di uova e di spermi aploidi con 21 elementi) durante le quali è documentata la ricombinazione.

Pertanto le artemie bisessuate danno luogo a una variabilità genetica, come qualsiasi altra specie, che trae origine dalla mutabilità spontanea e coinvolge naturalmente anche i caratteri di interesse pratico per la produzione in massa. Le artemie partenogenetiche costituiscono un gruppo eterogeneo (per quello che si riferisce al meccanismo di segregazione dei caratteri) anche perché, come Artom scoprì nel 1931, vi possono essere artemie partenogenetiche diploidi (con 42 cromosomi) oltre che tetraploidi (con 84) come si era creduto per molto tempo. Artemie diploidi partenogenetiche si trovano a Sète (Francia meridionale), a Margherita di Savoia (Puglia) e in altre località. Alcune partenogenetiche (le tetraploidi) sono praticamente prive di meiosi, cioè trasmettono il genotipo senza variazione da generazione a generazione, a meno di mutazioni insorte. Le artemie partenogenetiche diploidi presentano un complesso di varianti, durante la maturazione delle uova che possono permettere tanto la segregazione quanto la ricombinazione; anche queste forme, dunque, possono dar luogo a nuove combinazioni genotipiche come le bisessuate.

Queste considerazioni appaiono importanti se si considera un carattere che è diventato o diventerà importante nel prossimo futuro onde selezionare artemie per finalità pratiche: la grandezza cellulare. Questo carattere fu il primo a essere considerato come distintivo fra artemie con 42 e con 84 cromosomi; queste ultime hanno cellule maggiori. È un carattere che distingue quasi sempre le forme a numero cromosomico multiplo. Il carattere, scoperto da Artom fin dal 1912, venne sempre misurato sulle cellule dell'epitelio intestinale, cioè su elementi labili. Ciò che viene influenzato primariamente dal numero cromosomico è in realtà il nucleo: la dimensione totale della cellula ne è una conseguenza che risente di diverse influenze. Una di queste è un controllo dei geni, come venne da me

Fotografie a colori di artemie adulte. In alto a sinistra il capo e l'inizio del torace: si vedono le due seconde antenne a tenaglia che servono a trattenere la femmina durante l'accoppiamento. In alto a destra capo con antenne filiformi e inizio del torace di una femmina. In basso è fotografata una parte del corpo di una femmina con un gruppo di uova nell'ovisacco.

Fotografia di due larve (naupli) di artemia: a destra nauplius che esce dalla cisti (uovo ricoperto da un guscio e con l'embrione in stato avanzato); a sinistra nauplius in grado di nuotare.

Schemi di maturazione dell'uovo di artemia. A sinistra viene indicata una sola coppia di cromosomi, ciascuno formato da due bracci uguali e con un punto (lettera dell'alfabeto) corrispondente a un gene. Le lettere maiuscole e minuscole corrispondono rispettivamente alla versione dominante e a quella recessiva. A maturazione conclusa, uno solo dei cromosomi resta nell'uovo. Nelle combinazioni rappresentate (che sono solo una parte di quelle possibili) si vedono i due gameti prima di fondersi per formare lo zigote. Le diverse combinazioni spiegano come origini parte della variabilità genetica. Lo schema a destra si riferisce alla maturazione dell'uovo nell'artemia partenogenetica dove i cromosomi rimangono doppi.

La grandezza dei nuclei di un tessuto epiteliale monostratificato come quello che riveste internamente l'intestino può essere facilmente misurata come superficie massima del nucleo stesso, che è elissoidale. Partendo da singole femmine partenogenetiche dal numero cromosomico noto (84 in questo caso), la distribuzione delle grandezze entro la progenie di ogni femmina dimostra che esiste un controllo genetico che modifica la grandezza determinata dal numero cromosomico.

dimostrato nel 1940 facendo considerazioni statistiche sulle dimensioni cellulari (superficie delle cellule) dei ceppi allevati in laboratorio derivati ciascuno da una sola femmina partenogenetica tetraploide. Le differenze dimostrano chiaramente che, a parità di numero cromosomico, esiste un controllo trasmesso attraverso il genotipo, evidentemente risultato di mutazioni. Una questione molto importante, ma non ancora chiarita riguarda la dimensione cellulare con la sua complessa determinazione genica e cromosomica che probabilmente si riflette sulla dimensione dell'uovo embrionale durevole (cisti) e del nauplius. In effetti si sa già che esistono differenze dimensionali fra cisti di diverse popolazioni *(si veda la figura a pagina 89)* ed è urgente conoscere almeno il numero cromosomico ed eventuali variazioni della dimensione dei cromosomi in alcune delle più importanti popolazioni, tanto più che si sa che la dimensione delle cisti è direttamente proporzionale a quello del nauplius. Fortunatamente in questi ultimi mesi è stata messa a punto una nuova tecnica per l'analisi numerica, dimensionale ed eventualmente strutturale dei cromosomi del nauplius, così che il problema potrà essere risolto in breve tempo *(si vedano le figure a pagina 88).*

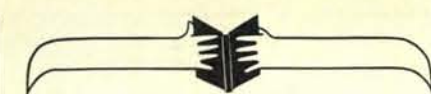
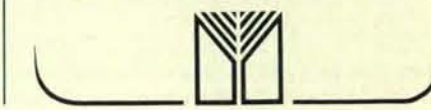
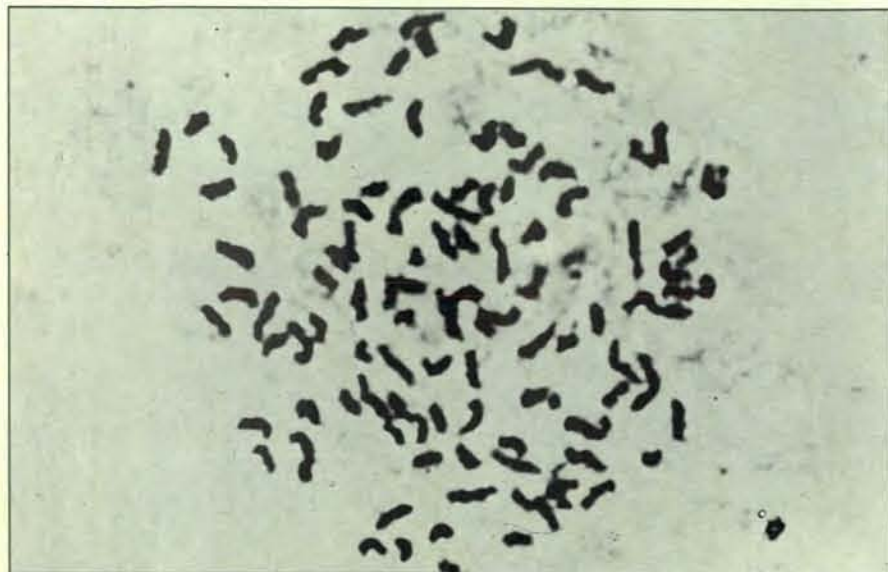
Un ultimo aspetto della biologia dell'artemia è quello del numero di specie che si ascrivono al genere *Artemia.* Nel 1968 fu definita con metodo sperimentale una seconda specie, oltre ad *A. salina,* denominata *A. persimilis.* È vero che Artom nel 1921 aveva tentato una nuova sistematica del genere distinguendo specie a nuclei grandi e a nuclei piccoli. Ma il tentativo non ebbe fortuna. Le altre specie descritte da A. E. Verrill fino dal 1869 rimasero entità di incerto significato, talvolta localizzate in aree che ne rendevano e ne rendono difficile lo studio. *A. persimilis* non differisce se non per caratteri morfologici minimi da *A. salina,* possiede 44 cromosomi invece di 42, più piccoli di quelli dell'altra specie, da cui è separata da una nettissima barriera genetica. Rapidamente ci si rese conto dell'esistenza di altre specie *(A. franciscana, A. tunisiana, A. urmiana,* ecc.) così che oggi il genere è un tipico esempio di speciazione basata sulla formazione di specie gemelle o criptiche. Alla tipizzazione di queste specie contribuisce molto e contribuirà ancora di più in futuro ogni ricerca di caratteri biochimici (polimorfismo della emoglobina, di enzimi e di altre proteine) come è stato fatto da S. T. Bowen e dal suo gruppo. Il significato di queste varianti mutazionali (come di altre trovate da me negli allevamenti di quella forma che allora era considerata sempre *A. salina)* agli effetti delle preferenze dell'utilizzazione industriale e di eventuali selezioni, non è stato ancora riconosciuto, mentre è probabile che le questioni tassonomiche (se conservare il nome *A. salina* o abolirlo, come classificare e denominare le forme partenogenetiche ecc.) debbano avere soprattutto significato teorico.

Come ho detto, l'interesse dell'industria per l'artemia emerse circa una ventina d'anni fa con la raccolta e vendita di cisti per ottenere larve vive per nutrire pesci ornamentali. Senza voler negare il valore di questa utilizzazione, essa è però di ben minore importanza rispetto a quello della nutrizione di crostacei (granchi o gamberi marini) e di pesci di interesse commerciale. Di questo nuovo fenomeno possiamo indicare come inizio un lavoro di P. Sorgeloos presentato a un convegno organizzato dalla FAO a Kyoto nel 1976, sebbene già tre anni prima Helfrich avesse considerato la possibilità di allevare artemie in massa alla Christmas Island nelle Hawaii.

In realtà l'allevamento dell'artemia iniziò come mezzo di studio della genetica di questa specie. I primi tentativi sistematici furono i miei a Milano nel 1939 ove si continuò fino al 1974, basando l'allevamento sulla nutrizione con alghe verdi vive *(Dunaliella, Stephanoptera)* per cui il presupposto per allevare e far riprodurre artemie era la produzione di alcuni litri al giorno di terreno con alghe vive in grande quantità. Nello stesso tempo, Bowen (dal 1962 in avanti) nutrì l'artemia con lievito, ottenendo buoni risultati e una notevole semplificazione tecnica. Da quando furono imposte finalità pratiche all'allevamento dell'artemia il problema ha trovato soluzioni del tutto nuove.

Dobbiamo ora esaminare in particolare queste finalità pratiche, utilizzando l'esperienza ancora unica compiuta da Sorgeloos nel 1978 e 1979, basata sulla collaborazione ai programmi in atto in Tailandia e nelle Filippine che possono servire di esempio e di modello per altri programmi già in via di realizzazione o futuri.

L'utilizzazione dell'artemia in generale può essere così riassunta: *1)* produzione delle cisti di larve (nauplius), di forme in via di sviluppo o di adulti per alimentare crostacei e pesci; *2)* produzione di proteine da alimentazione vegetale (come si è visto l'alimentazione naturale dell'artemia è costituita dalle alghe verdi) da usarsi in diverse tecnologie; *3)* produzione di cisti da mettere in vendita.

Le specie che finora vengono vantaggiosamente nutrite con artemia sono: larve di crostacei come *Penaeus monodon, Macrobrachium rosenbergii, Scylla serrata* e *Chaos chaos* e alcune specie di pesci. Le artemie già sviluppate o adulte possono essere impiegate con vantaggio specialmente per nutrire crostacei e pesci adulti. Si deduce dunque che il problema dell'utilizzazione industriale dell'artemia si riduce essenzialmente alla produzione di cisti (che possono essere prodotte da chi alleva crostacei e pesci o possono essere acquistate) di cui vanno determinate

Fase di mitosi preparata con la nuova tecnica messa a punto da C. Barigozzi e L. Baratelli in una cellula somatica (nauplius). I cromosomi sono 42 e corrispondono al normale stato diploide.

Fase corrispondente a quella della figura in alto e preparata con la medesima tecnica, ma appartenente a un individuo tetraploide. L'ingrandimento in entrambe le foto è di 1800 diametri.

STATO	VARIETÀ DI ARTEMIA
BRASILE	CEPPO DI MACAU (BAIA DI SAN FRANCISCO)
FRANCIA	CEPPO DI AIGUES MORTES
SPAGNA	CEPPI DELLA ZONA DI CADICE
STATI UNITI	CEPPO DELL'ARGENTINA
	CEPPO DI SHARK BAY (AUSTRALIA)
	CEPPO DELL'AUSTRALIA OCCIDENTALE
	CEPPO CANADESE (CHAPLIN LAKE)
	CEPPO DEL GRAN LAGO SALATO
	CEPPO DI SAN DIEGO
	CEPPO DELLA PARTE EST DELLA BAIA DI SAN FRANCISCO
	CEPPI DEL NORD E DEL SUD DELLA BAIA DI SAN FRANCISCO

Molte ditte degli stati sopra elencati diffondono listini indicanti la provenienza delle cisti, il che rappresenta tuttora il dato informativo più facile, anche se il meno accurato, per classificare il materiale.

principalmente la schiudibilità, la durata del periodo richiesto per schiudere e la dimensione ottimale in relazione alle dimensioni del nauplius. Va detto che le cisti possono essere conservate anche a lungo, purché vengano protette dall'aria, in quanto l'ossigeno atmosferico ne riduce la schiudibilità.

Prima di discutere il problema delle tecniche per ottenere le cisti in massa è necessario dare un'idea della quantità di questo materiale da usarsi e del relativo costo.

Le cisti si vendono a chilogrammi e i naupli si valutano a milioni; la produzione attuale nel mondo viene stimata in oltre 100 000 chilogrammi ed è prevedibile un suo forte aumento annuale. Le cisti si vendono al prezzo di circa 50 dollari al chilogrammo. Siccome l'ordine di grandezza della quantità per iniziare un programma di produzione è di 150 chilogrammi, la spesa per l'acquisto di cisti può raggiungere i 7-8 mila dollari.

Vediamo ora come si possono ottenere tali quantità di cisti. Due sono i metodi oggi a disposizione, oltre alla raccolta di cisti deposte naturalmente nelle saline: l'introduzione di cisti in bacini salati, nei quali vi trovino nutrimento appropriato, come per esempio l'alga *Spirulina,* che può essere ottenuta secca dal Messico in grandi quantità e trasportata ove si vuole, e l'allevamento in condizioni artificiali.

Per l'allevamento in bacini naturali si immettono le cisti in acqua; per esempio, da 50 grammi di cisti provenienti dal Gran Lago Salato dello Utah si ottiene in 36 ore a 28° 1 milione di naupli. Nei bacini salati di Macau in Brasile si ottennero i seguenti risultati: nell'aprile del 1977 vennero inoculati 250 grammi di cisti di artemia provenienti dalla Baia di San Francisco; già nel giugno 1977 avvenne la raccolta dei primi chilogrammi di cisti da alcune migliaia di ettari di bacino e nel dicembre 1977 si ottenne la raccolta di 3 mila chilogrammi di cisti di buona qualità.

Questi valori danno un'idea della cifra di partenza. Un procedimento consigliato da Sorgeloos per migliorare la schiusura delle cisti (che, come vedremo, è un punto di grande importanza) è la loro decapsulazione in massa dopo trattamento con ipoclorito di calcio e successiva conservazione in acqua salata a temperatura ambiente. Anche dopo un certo tempo danno origine a naupli perfettamente vitali.

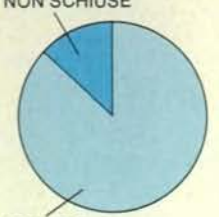
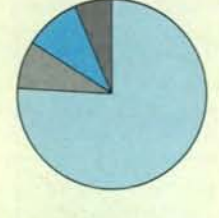
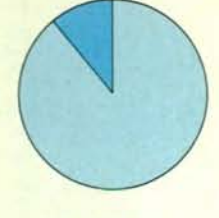
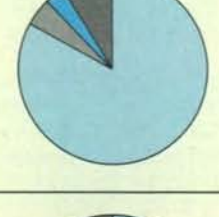
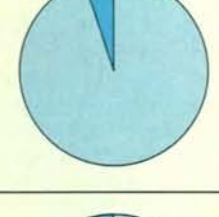
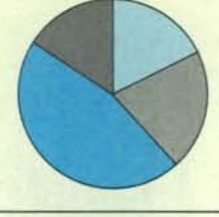
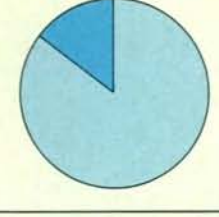
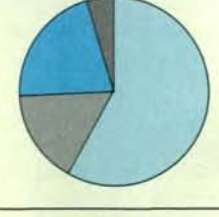
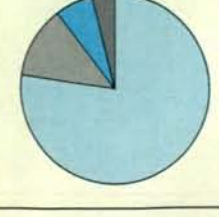
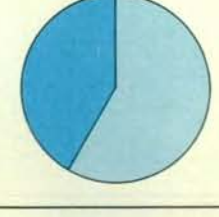
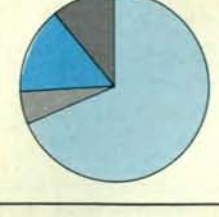
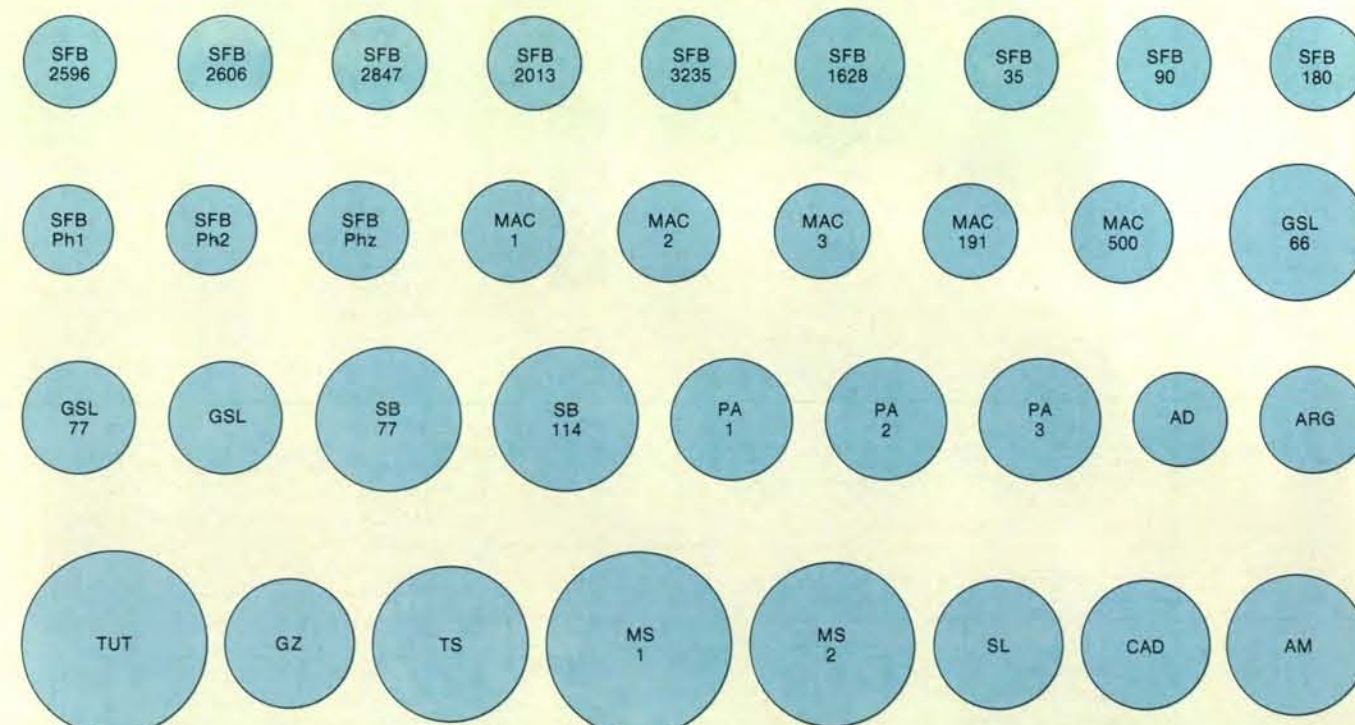
Questo allevamento, che si avvicina alle condizioni naturali, dovrebbe comprendere anche recipienti ove ottenere forme più sviluppate di nauplius e individui adulti, specialmente se lo scopo dell'allevamento non è solamente la nutrizione di crostacei e pesci, ma anche l'ottenimento in larga scala di proteine che costituiscono circa il 60 per cento del corpo di una artemia adulta e che si sono dimostrate di ottima qualità.

Per l'allevamento in condizioni artificiali la tecnica più interessante è quella elaborata da Sorgeloos che si basa sull'utilizzazione di un alimento inerte mai usato precedentemente: la pula di riso polverizzata, che è un sottoprodotto di bassissimo costo. Finora un impianto pilota si trova presso il laboratorio di maricoltura dell'Università di Gand (Belgio). L'aspetto più interessante di questa tecnica è questo: la pula di riso (ridotta a frammenti inferiori ai 60 micrometri) ha un contenuto in amminoacidi assai basso (10,51 per cento del peso secco) mentre le artemie nutrite con questo alimento hanno un contenuto del 51,14 (sempre in percentuale sul peso secco). Ciò vuol dire che l'artemia è un potentissimo mezzo per produrre naturalmente proteine. Se si pensa che le proteine dell'artemia potrebbero servire ai più diversi impieghi nell'alimentazione (nella mangimistica, per esempio), si può facilmente concludere che questo piccolo crostaceo può contribuire alla soluzione di problemi di vasta portata.

La strutturazione dell'impianto d'allevamento in laboratorio di artemia è piuttosto complessa. Essa deve tenere conto, anzitutto, della necessità di un afflusso continuo di alimento: ogni rallentamento o interruzione dell'afflusso di cibo può essere dannoso. Perciò esiste un dispositivo che regola automaticamente l'afflusso di polvere di pula di riso, non appena la concentrazione di questa diventi troppo bassa. Altri dispositivi raccolgono le cisti e allontanano le feci. Una descrizione più completa dell'impianto non è necessaria, anche perché è probabile che miglioramenti futuri modifichino sensibilmente la struttura attuale, ancora in fase di sperimentazione.

La produzione industriale di cisti di artemia ha già raggiunto un tale sviluppo che in diversi paesi del mondo esiste ormai una rete di distributori ben organizzata. Le specie diffuse commercialmente sono elencate nella tabella in basso nella pagina a fronte.

Va però detto che il materiale messo in vendita non è affatto omogeneo e ciò si riflette in una diversa efficienza per l'allevamento. Infatti i valori che hanno valore economico prevalente sono: *1)* la proporzione di schiusa delle cisti che si riflette sul numero dei naupli prodotto da una data quantità di cisti; *2)* il tempo richiesto per la schiusura che può essere un efficace indicatore della velocità di sviluppo; *3)* la resistenza alla decapsulazione delle cisti che, in genere, sembra essere molto elevata; *4)* il rapporto fra dimensione delle cisti e dimensione del nauplius che è ancora da studiare, ma che sicuramente rivela la correlazione esistente tra dimensione cellulare e numero cromosomico.

Tutti questi fattori sono indubbiamente controllati dal genotipo e rivelano una eterogeneità genetica fra i ceppi commerciali ed anche al loro interno che è ancora quasi completamente inesplorata.

Ho già indicato la variabilità delle dimensioni delle cisti di diverse popolazioni. Basterebbero questi dati (anche corredati dal numero cromosomico che è già una indicazione della grandezza cellulare del nauplius, delle forme di sviluppo dell'adulto) per affermare la necessità di ricerche biologiche (genetiche, biochimiche, ecc.) per poter indicare se convenga prescegliere certe popolazioni anziché altre, allo scopo di raggiungere una certa dimensione media almeno del nauplius ed, eventualmente, iniziare un lavoro di selezione. A questa necessaria conoscenza, si deve aggiungere quella della schiudibilità delle cisti, che si indica come quantità di cisti in grammi necessaria per ottenere un milione di naupli. Sorgeloos ha messo a confronto cinque popolazioni di diversa provenienza considerando il tempo loro necessario per la schiusura: due popolazioni della Baia di San Francisco, quella del Gran Lago Salato dello Utah, quella di Shark Bay in Australia e quella di Buenos Aires. Le differenze sono grandissime (le cisti del Gran Lago Salato hanno una rapidità di schiusura minima) onde si pone il problema della selezione di ceppi a schiusura massima ricavate dalla popolazione più promettente, come alcune della Baia di San Francisco *(A. franciscana).* Questo lavoro potrebbe essere inutile solo se risultasse come fatto certo che la decapsulazione delle cisti di ogni provenienza possa elevare così fortemente la produzione di naupli da annullare o quasi le differenze fra popolazioni. Per renderci conto di quanto sia importante il tasso di schiusura basterà ricordare che, secondo i dati forniti da Sorgeloos, 50 grammi di cisti del Gran Lago Salato sono necessari per ottenere un milione di naupli, mentre con le artemie di San Francisco (Leslie Salt Pond, San Francisco Bay) bastano 5,5 grammi per ottenere la stessa quantità di naupli. Un altro importante elemento è la durata dell'intervallo fra il momento nel quale si pongono le cisti in acqua salata e il momento della schiusa: per le artemie del Gran Lago Salato 36 ore a 28°, 24 ore a 28° per quelle di San Francisco. Può

La dimensione della cisti è uno dei caratteri potenzialmente più importanti per la scelta dei ceppi da allevamento. La figura illustra la varietà dimensionale finora riscontrata in popolazioni di diversa provenienza geografica. In generale si nota una correlazione diretta fra grandezza delle cisti e numero cromosomico (le più grandi appartengono a popolazioni poliploidi). Le sigle iscritte nei cerchi indicano la località di provenienza: SFB = San Francisco Bay; MAC = Macau (Brasile); GSL = Great Salt Lake (USA); SB = Shark Bay (Australia); PA = Port Araya (Venezuela); AD = Adelaide (Australia); ARG = Argentina; TUT = Tuticorin (India); GZ = Galera Zamba (Colombia); TS = Tien Tsin (Cina); MS = Margherita di Savoia (Italia); SL = Sanlúcar (Spagna); CAD = Cadice (Spagna); AM = Aigues Mortes (Francia).

Rappresentazione schematica della efficienza di schiusa di campioni di cisti ricavate da cinque tipi commerciali di artemia. La valutazione è stata fatta in base agli elementi che compongono le cisti: *1)* cisti intera, *2)* rivestimenti leggeri (guscio), *3)* rivestimenti pesanti (sabbia ecc.), *4)* acqua.

essere una differenza interspecifica, ma può essere anche risultato di particolare genotipi presenti in una certa percentuale in una data popolazione, che potrebbero presentarsi a un miglioramento per selezione.

È ovvio, dopo quanto si è detto, che la selezione può essere pensabile solo nelle forme bisessuate e in certe forme partenogenetiche diploidi; non sarebbe prevedibile in quelle partenogenetiche tetraploidi che sono ameiotiche.

Esistono però ancora molti aspetti inesplorati della biologia nell'artemia che certamente vanno considerati sotto il profilo genetico: il rapporto fra salinità e produttività, la qualità delle proteine sintetizzate, l'oviparità rispetto alla viviparità (cioè rispetto alla produzione di naupli, senza passare attraverso lo stadio di cisti) e la resistenza alle malattie.

Come si vede sono i problemi che (sia pure sotto diverso aspetto) si presentano in tutte le specie allevate dall'uomo in stato artificiale per raggiungere finalità diverse da quelle prodotte dalla selezione naturale.

Questo stato di cose ci mette di fronte a una situazione che impone una strategia molto ragionata, che non può essere concepita se non su scala internazionale. Da tempo si sa che vengono trasportate artemie da un biotopo all'altro, ciò che tende a modificare drasticamente la distribuzione geografica dei componenti del genere *Artemia*. Le conoscenze incomplete sul controllo genotipico di caratteri economicamente importanti non danno alcuna garanzia che i campioni introdotti siano i migliori dal punto di vista economico. Vi possono essere, anzi, casi in cui una buona e piccola popolazione locale venga eliminata o si incroci con le artemie introdotte. Questa seconda possibilità potrebbe anche dare risultati validi, ma ciò dovrebbe essere accuratamente controllato. Ci troviamo, per analogia, di fronte a un problema simile a quello posto dalle piante di interesse agricolo, qualora si diffondessero varietà non sufficientemente note e controllate.

D'altro canto esistono numerosissime popolazioni di artemia (forse anche corrispondenti a specie diverse) tuttora sconosciute. Solo in Italia vi sono almeno quattro o cinque popolazioni tuttora da studiare; in Israele o in Giordania dovrebbero esservi artemie altamente poliploidi che nessuno ha esaminato da tempo. Nell'Unione Sovietica, probabilmente a Odessa, vi potrebbero essere artemie bisessuate tetraploidi, la cui esistenza è sempre dubbia.

Data l'importanza assunta dall'artemia, queste popolazioni vanno considerate alla stessa stregua delle varietà selvatiche delle piante coltivate che si conservano e si studiano nelle banche dei geni o nei laboratori del germoplasma.

Credo sia venuto il momento di pensare a pianificare uno studio e una conservazione di ceppi di artemia, allo scopo di permettere quelle selezioni (o eventualmente quegli incroci) che possano venire incontro alle necessità della produzione.

Antiche navi da guerra a remi

Le città-stato della Grecia classica usavano flotte di navi munite di sperone che avevano sino a tre banchi di remi. Esse si trasformarono in seguito in enormi piattaforme armate con migliaia di rematori

di Vernard Foley e Werner Soedel

Durante i periodi aurei della Grecia e di Roma navi da guerra con propulsione prevalentemente a remi svolsero un ruolo determinante nel rendere possibile il commercio su lunghe distanze e nell'assicurare la sicurezza dell'impero. La vela, che era stata introdotta attorno al 3500 a.C., fu adattata primariamente alle navi mercantili, poiché i costi della mano d'opera imponevano di contenere quanto più possibile in tale campo le dimensioni degli equipaggi. Le navi mercantili facevano uso dei remi solo per salpare e per entrare in porto o per non restare immobilizzate in mezzo al mare durante bonacce prolungate. Anche le navi da guerra facevano affidamento il più possibile sulle vele, ma nell'imminenza della battaglia le vele principali venivano ammainate e depositate col relativo sartiame su qualche spiaggia conveniente. I remi fornivano meno energia di quanta ne somministrasse un buon vento, ma per brevi periodi conferivano a una nave da guerra una velocità adeguata e una superiore manovrabilità. Il metabolismo dei rematori costituiva un modo per immagazzinare energia in quantità significativa e per liberarla poi rapidamente, un vantaggio inaccessibile alla vela.

Nonostante la relativa inefficienza della propulsione a remi per viaggi su lunghe distanze, le antiche navi da guerra a remi conseguirono un livello impressionante di sofisticazione tecnica. La nave da guerra standard del periodo in cui si esercitò l'influenza della marineria greca fu la trireme. Il vocabolo deriva dal latino *triremis*, che proviene a sua volta dal greco *trieres*. Un tempo ci furono grandi controversie fra i classicisti circa l'esatta interpretazione del vocabolo greco, che significa approssimativamente «equipaggiato per tre». Oggi il problema è risolto in modo abbastanza soddisfacente, grazie in gran parte ai lavori dello studioso britannico J. S. Morrison, che è riuscito a ricostruire un'immagine attendibile di tali famose navi sulla scorta di una varietà di dati, come citazioni letterarie, iscrizioni in cantieri, documenti iconografici che ci sono pervenuti e misurazioni degli scali in cui venivano tirate in secco le navi per procedere a lavori di manutenzione e a piccole riparazioni.

La trireme si era sviluppata da tipi più semplici principalmente greci e fenici. Queste navi primitive erano scafi aperti la cui propulsione era assicurata da una fila di rematori su ciascun lato dello scafo stesso. Dalla pittura vascolare e da alcuni passi di Omero risulta che erano costruite in modo da raggiungere una velocità elevata e da utilizzare in modo efficiente l'energia degli equipaggi, ma per consentire loro di raggiungere un'efficacia massima si richiedeva ancora un altro elemento. Quest'innovazione coincise con la comparsa, attorno all'800 a.C., dello sperone, che determinò una rivoluzione nell'antica architettura navale. In precedenza le battaglie navali si risolvevano in definitiva in uno scontro uomo contro uomo, che si realizzava attraverso l'abbordaggio. Ora si poté mirare invece all'affondamento della nave avversaria, anziché a infliggere solo perdite agli equipaggi nemici.

L'avvento dello sperone rese molto più sentita l'esigenza di una grande velocità e manovrabilità. Perciò le semplici navi dei tempi preomerici si trasformarono ben presto in navi dallo scafo lungo, basso e sottile, che poteva alloggiare sino a 25 rematori su ciascun lato. Queste navi furono chiamate pentecòntori, o navi a cinquanta remi. Le dimensioni del loro scafo, che sono state determinate in modo abbastanza certo, rivelano vari princìpi di protezione che si conservarono nel futuro sviluppo di navi più elaborate.

La resistenza di una nave al moto attraverso l'acqua è governata da quattro fattori principali. Uno di essi è la resistenza d'attrito, la quale è una conseguenza del fatto che l'acqua, come tutti gli altri fluidi, presenta una certa viscosità. Un altro fattore è la resistenza dovuta alla figura, ossia al grado di idrodinamicità del suo profilo. Quando le molecole d'acqua non possono fluire a contatto della chiglia in linee uniformi e ininterrotte, si separano dallo scafo, accrescendo in tal modo il volume dell'acqua che dev'essere spostata. Nel caso che se ne distacchino in misura sufficiente, formano vortici. L'energia che si richiede per formare i vortici viene sottratta alla quantità di energia disponibile per muovere la nave, cosicché la resistenza creata dai vortici è una terza forma di ostacolo che si frappone al movimento della nave stessa.

Questi tre fattori sono tutti strettamente interconnessi. Un quarto genere di resistenza, quello della produzione di onde, può essere trattato in modo relativamente indipendente. Come gli altri fattori, anch'esso aumenta al crescere della velocità della nave, ma il suo aumento è tanto più rapido di quello degli altri che, al raggiungimento di certe velocità, esso diventa il principale fattore di resistenza. Esso è determinato soprattutto dal rapporto fra la lunghezza della nave e la lunghezza delle onde create dal moto di questa. Per comprendere la ragione di questo stato di cose, occorre visualizzare il modo in cui le onde prodotte da una nave interagiscono col suo scafo al crescere della velocità.

Quando la prua viene a contatto con nuova acqua, impartisce a una parte di essa un'accelerazione verso l'alto che varia con la velocità. La gravità svolge un'azione antagonistica nei confronti di questo moto e infine l'acqua ridiscende al livello della superficie e al di sotto di esso. Così si forma un'onda. Nel frattempo la nave si è spostata in avanti di un piccolo tratto. L'effetto combinato dei due moti è quello di determinare la formazione di una o più onde che sono stazionarie rispetto allo scafo quando la velocità della nave è costante. A basse velocità ci sarà un certo numero di creste d'onda, poiché le singole molecole d'acqua avranno il tempo di salire e ridiscendere varie volte prima che la nave sia passata.

All'aumento di pressione a prua, che determina la formazione di queste onde, corrisponde una diminuzione di pressione a poppa, che inizialmente tende ad abbassare la superficie dell'acqua. Entrambi questi insiemi di disturbi trasferiscono energia dalla nave all'acqua e aumentano ulteriormente la resistenza dell'acqua che si esercita sullo scafo, aumentandone la superficie che entra a contatto con l'acqua. Se la velocità dello scafo è tale che l'onda stazionaria creata dalla prua è in fase con la depressione dell'onda a poppa, queste differenze di livello nell'acqua create dalla pressione si rafforzano, mentre se l'onda di prua e la depressione di poppa sono completamente fuori fase praticamente si annullano. Quanto maggiore è la velocità dello scafo tanto più aumenta la lunghezza d'onda del sistema di disturbo, cosicché una nave che si acceleri a partire dallo stato di quiete passa per successive regioni di velocità in cui la resistenza cresce con rapidità in un primo tempo maggiore e poi minore. In altri termini, la curva della resistenza sale in un certo senso a zig-zag.

Infine si determina una differenza di immersione fra prua e poppa. La poppa si abbassa e la prua si solleva, cosicché la nave deve tentare di scalare una collina d'acqua che ha creato essa stessa. Quest'effetto diventa critico quando la lunghezza d'onda dell'onda stazionaria di prua raggiunge la lunghezza dello scafo. A questo punto per poter aumentare ulteriormente la velocità della nave è necessario impartirle una potenza molto maggiore di quella fornitale sino a questo momento. Ne consegue che, quanto maggiore è la lunghezza dello scafo, tanto più a lungo è possibile rimandare questa crisi nella velocità. Uno scafo lungo e sottile minimizza quindi i disturbi di prua e di poppa assorbendo una quantità maggiore della pressione totale necessaria a sostenere la nave lontano da questi due punti e distribuendola invece lungo i fianchi, dove contribuisce con minore rapidità, attraverso l'attrito superficiale, all'aumento totale della resistenza.

Il limite cruciale alla velocità massima di una nave che abbia una sorgente di energia fissa è dato dalla relazione di Froude: il rapporto tra la lunghezza dello scafo e il quadrato della velocità (dal momento che la resistenza tende ad aumentare col quadrato della velocità). Questa relazione prende il nome da William Froude, l'ingegnere navale britannico che per primo chiarì questi problemi negli anni settanta del secolo scorso.

Ovviamente non ci si può attendere che quest'analisi sia già stata compiuta dagli antichi. Provando e riprovando, essi pervennero però a una comprensione efficace dei principali fattori che impedivano a una nave di acquisire una velocità elevata. Lo scafo di una pentecòntoro era lungo sino a 38 metri, e non era largo probabilmente più di quattro. Il rapporto approssimativo di 10:1 fra lunghezza e larghezza tipico delle poliremi progettate per raggiungere le massime velocità si conservò sino alla fine dell'era delle navi da guerra a remi nell'antichità. Gli specialisti sono oggi generalmente d'accordo sulla tesi che le lunghezze raggiunte da tali navi erano molto vicine ai valori massimi raggiungibili con costruzioni in legno. In effetti pare che, nel caso delle triremi ancor più sottili, i limiti intrinseci alla costruzione in legno siano stati addirittura superati. Anche se un complesso sistema di mortase, tenoni e caviglie univa il fasciame della nave in modo tale che le sollecitazioni venivano distribuite in misura considerevole sull'intero scafo, non era prudente far prendere il mare a una trireme che non fosse stata rafforzata da grandi cavi che la cingevano da prua a poppa e che erano stati sottoposti a una tensione considerevole da un argano. Non sappiamo dove fossero situati esattamente tali cavi, ma pare che l'effetto di compressione che essi esercitavano sulla nave fosse necessario per impedire allo scafo di indebolirsi pericolosamente quando veniva sottoposto a sollecitazioni. Il legno di per sé è poco idoneo ad adattarsi a giunti soggetti a tensione.

La larghezza di questi scafi era poco più

Il migliore documento iconografico dell'aspetto della trireme greca alle sue origini è questo frammento di un rilievo in pietra scolpito verso la fine del V secolo a.C. Noto come rilievo Lenormant dal nome dell'archeologo francese Charles Lenormant che lo scoprì nel 1852, si trova oggi nel museo dell'Acropoli ad Atene. Esso fu eseguito presumibilmente in scala. Come nell'illustrazione di copertina, sono visibili solo i vogatori dell'ordine superiore; gli altri due banchi sono più in basso nello scafo o si trovano all'interno rispetto ai vogatori del banco superiore.

del minimo sufficiente a far posto a due rematori seduti a fianco a fianco i cui remi poggiavano sul parapetto (il bordo della murata). Il punto di appoggio del remo (la resistenza o il fulcro secondo che si veda nel remo una leva di 2° genere o di 1° genere) non può essere troppo vicino all'impugnatura, altrimenti lo sforzo del rematore diventa eccessivo. Nell'uso moderno il remo viene mantenuto per un terzo (o poco meno) della sua lunghezza all'interno del bordo dell'imbarcazione. Poiché i remi delle triremi avevano una lunghezza comune ancor oggi nei remi, si può supporre che lo stesso rapporto fosse utilizzato nell'antichità. Considerando infine che fra le estremità dei remi è necessario un po' di spazio libero per consentire i movimenti delle spalle dei rematori adiacenti, risulta evidente che lo scafo di una pentecòntoro racchiudeva le sue unità di propulsione entro uno spazio dalla larghezza veramente esigua. Anche il pescaggio dello scafo era minimo, forse di mezzo metro, cosicché corrispondentemente modesto era il volume del liquido spostato e minima la resistenza d'attrito.

Gli scafi di queste navi potevano avere un pescaggio modesto, essendo formati di legno tenero, leggero, usato inoltre con parsimonia. Lo scafo aveva uno spessore di soli 3,5 centimetri circa e in taluni punti era ancor più sottile. Parlando dello scafo delle navi mercantili, che di solito erano costruite in modo più massiccio, un antico poeta osservò che solo tre dita di legno separavano il marinaio da una sorte tragica. Pare che la riduzione del peso dello scafo sia stata spinta a un punto tale che l'equipaggio veniva a costituire un terzo circa della massa totale del sistema. Noi stimiamo che il dislocamento della trireme fosse inferiore a 40 tonnellate compresi i rematori.

La configurazione particolareggiata dello scafo dava altri contributi alla sua velocità. Oggi ci è familiare la linea delle superpetroliere, con la prua a bulbo proiettata in avanti sott'acqua per minimizzare la formazione di un'onda di prua, evitando improvvise variazioni della quantità di moto nell'acqua. Gli antichi speroni presentavano una varia configurazione, ma almeno alcuni di essi dovevano avere lo stesso effetto. A poppa le poliremi si sollevavano in modo graduale e regolare dall'acqua come qualsiasi moderno yacht da regata, riducendo in tal modo al minimo la formazione di vortici e di una scia. In effetti, le poliremi potevano avere una linea più dura delle imbarcazioni moderne, le quali devono spesso incorporare un piano verticale per sostenere il timone. Le navi dell'antichità venivano guidate per mezzo di due remi ai due lati della poppa. Questi remi bloccavano probabilmente la formazione di vortici e al tempo stesso mantenevano a valori minimi l'area dello scafo bagnata.

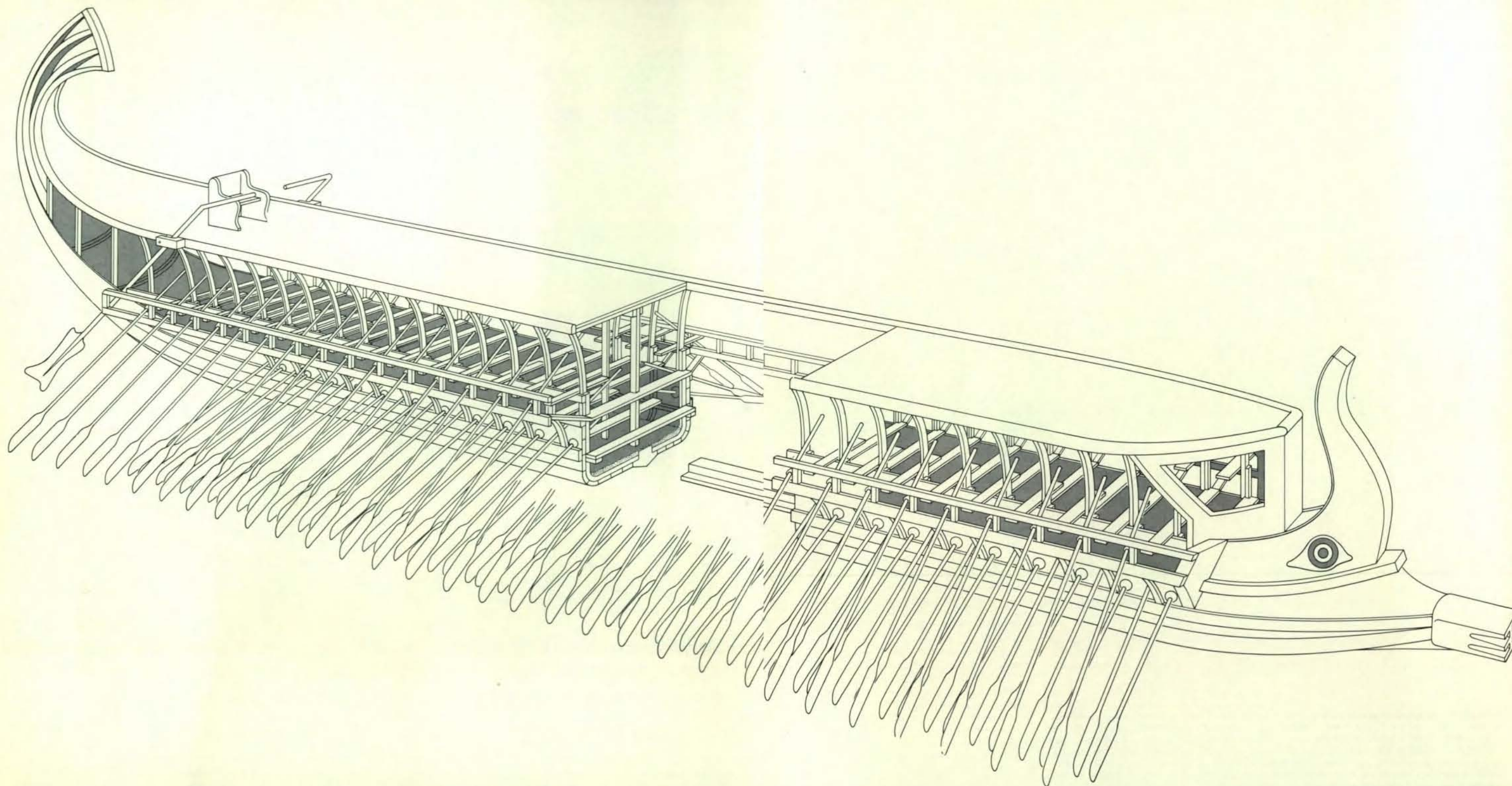
Al di sopra dell'acqua queste imbarcazioni presentavano un'uguale razionalità di progettazione: lo scafo basso minimizzava la resistenza del vento, tranne che a poppa, dove la curva ascendente dello scafo era prolungata nel cosiddetto aplu-

stre. Questo particolare costruttivo non solo contribuiva a proteggere lo scafo basso da ondate di poppa, ma, aumentando il momento dell'inerzia di rotazione dello scafo, contribuiva anche a diminuire il rollio, che è particolarmente sensibile in navi poco zavorrate. L'aplustre - agendo come una sorta di vela - contribuiva anche a orientare la nave secondo il vento quando veniva investita da una forte raffica, riducendo così il rischio per la nave di essere inondata da onde laterali.

Ancor prima dell'età delle triremi, quindi, le navi a remi erano pervenute a incorporare un livello di esperienza nautica considerevole. Questo si rifletteva nella loro velocità: la pentecòntoro più veloce aveva una velocità massima stimata a 9,5 nodi (17,6 chilometri all'ora), che è inferiore di solo un nodo circa alle migliori prestazioni di una moderna imbarcazione da regata.

Oltre al loro aspetto funzionale, le navi a remi avevano un aspetto emotivo che pare inscindibile dai più grandi tipi di navi della storia. Lo scafo era impermeabilizzato con pece e fori di ventilazione o chiazze di colore contrastante dipinti sulla prua immediatamente al di sopra e dietro lo sperone che poi si evolsero nelle cubie che alcune navi ostentano ancor oggi. Lo sperone era spesso rivestito di bronzo e decorato con disegni multipli di pugnali o conformato come il muso di un gigantesco cinghiale. Così una polireme che si stava avvicinando per speronare appariva nera, con uno sguardo fisso, e presentava a poppa una coda sollevata che la faceva assomigliare a un animale feroce. Talvolta nella parte superiore delle fiancate le poliremi presentavano delle protezioni contro le onde formate da pelli di animali conciate alle quali era stato lasciato il pelo. I remi accentuavano l'effetto zoomorfo; il lampeggiamento sincronizzato delle pale imbiancate dal sale che si sollevavano e si abbassavano era spesso paragonato dagli antichi poeti al movimento delle ali di un uccello. I remi imitavano persino l'iridescenza di certi animali; Aristotele osservò che in certe condizioni di luce, quando i remi frangevano l'acqua, i loro spruzzi formavano arcobaleni.

I costruttori erano però motivati principalmente dall'aspetto funzionale delle navi a remi e, fra il tempo di Omero e il 500 a.C. circa, furono compiuti vari passi avanti fondamentali in uno sforzo per racchiudere una maggiore potenza dinamica nello scafo della pentecòntoro. Non tutti i particolari sono ancora stati chiariti e l'intera storia di questi sviluppi è troppo complessa per poter essere presentata in breve, ma sostanzialmente si può dire che gli scafi furono prolungati verso l'alto mediante l'aggiunta di ponti o banchi sovrapposti per poter ospitare altri vogatori. Un secondo ordine di rematori ebbe come conseguenza la formazione di una bireme, e l'aggiunta di un terzo ordine dette origine alla trireme.

Per poter compiere questi passi avanti

Spaccato di una trireme greca del V secolo a.C.: si noti come i tre banchi dei vogatori fossero sistemati a scalinata in modo da consentire una disposizione il più possibile compatta. I 170 vogatori erano distribuiti nel modo seguente: 31 in ciascun banco superiore e 27 in ciascuno dei banchi inferiori. Poiché verso poppa lo scafo andava restringendosi, ivi non c'erano vogatori nei banchi inferiori. La nave era lunga complessivamente 35 metri circa e larga 3,5. La parte frontale dello scafo si prolungava in uno sperone, la principale arma offensiva della nave. La guida avveniva per mezzo di una coppia di remi speciali appesi fuori delle murate a poppa. Questa ricostruzione si fonda sull'opera dello studioso britannico J. S. Morrison, che a sua volta attinse massicciamente al rilievo di Lenormant, mostrato nella fotografia a pagina 95.

senza pregiudicare la stabilità dello scafo, che era già minima, si dovette procedere con cura. Già nel caso della bireme, se il secondo ordine di vogatori fosse stato collocato semplicemente al di sopra del primo, la nave avrebbe potuto avere un baricentro spostato pericolosamente verso l'alto. Nel progetto finale si ebbe invece solo una piccola estensione verso l'alto dei fianchi della nave, forse meno di mezzo metro. Al bordo di questo scafo remava come già nella pentecòntoro, una fila di uomini, ma ora essi costituivano solo lo strato di mezzo di un gruppo di tre. Al di sotto di essi, nella stiva della nave, vogavano uomini che erano vicini alla superficie del mare - mezzo metro circa - perché si potesse correre il rischio di abbassare il bordo dello scafo sino a quel livello. Perciò i loro remi fuoriuscivano da portelli praticati nella fiancata della nave, e guarnizioni di pelle assicurate al bordo del portello e al ginocchio del remo impedivano all'acqua di entrare nella nave.

Al livello superiore della nave, dove si trovava un terzo ordine di rematori, c'era un telaio a sbalzo o posticcio, un'innovazione che rese possibile il passaggio dalla bireme alla trireme. Il posticcio, sporgente fuori dalle murate, consentì ai maestri d'ascia di spostare il fulcro dei remi superiori a una sessantina di centimetri circa fuori del bordo della nave. I vogatori del banco superiore poterono sedere, grazie a ciò, all'altezza della spalla dei rematori dell'ordine di mezzo anziché al di sopra della loro testa, riducendo così di forse 50 centimetri quella che sarebbe stata senza questo accorgimento la crescita in altezza dello scafo.

La fisica della stabilità delle navi fu studiata per la prima volta in modo sistematico da Archimede, che visse nel III secolo a.C. Egli concepì il peso della nave come concentrato nel suo centro di gravità e diretto verso il basso verso il centro della Terra *(freccia nera)*. La forza di galleggiamento, diretta verso l'alto *(freccia colorata)*, era al centro di massa della parte dello scafo immersa *(in alto)*. Quando la nave rollava, queste due forze la facevano tornare alla verticale, perché il centro di massa della parte immersa si muoveva nella direzione del rollio *(al centro)*. In termini nautici moderni, si dice che il centro di galleggiamento si trova a una certa distanza verticale dalla linea mediana della nave; la distanza verticale fra il centro di massa della nave e l'intersezione della linea mediana della nave con una linea proiettata verso l'alto a partire dal centro di galleggiamento è l'altezza metacentrica *(in basso)*. Quanto maggiore è quest'altezza, tanto più stabile è la nave.

La sistemazione compatta dei rematori all'interno dello scafo fu resa ancora più efficace attraverso una disposizione a scalinata, per cui gli uomini dell'ordine superiore erano davanti di circa mezzo metro rispetto a quelli dell'ordine di mezzo e questi si trovavano spostati più in alto e più in avanti press'a poco della stessa misura rispetto agli uomini alloggiati nella stiva. Anche all'interno di ciascun livello gli uomini erano disposti molto vicini l'uno all'altro, essendo separati fra loro di non più di un metro circa. Perciò, se un uomo avesse remato anche solo leggermente fuori tempo, la sua schiena sarebbe venuta a contatto con le nocche dell'uomo seduto dietro di lui, e le nocche delle sue mani sarebbero venute a contatto col dorso dell'uomo seduto davanti a lui.

Gli uomini che vogavano nella stiva avevano anche un altro problema, connesso anche questo alla scarsità di spazio. Come osservò il commediografo Aristofane nella commedia *Le rane*, quando si spingevano in avanti per cominciare la vogata il loro naso arrivava molto vicino al fondo della schiena degli uomini seduti immediatamente di sopra e davanti a loro. Nello sforzo della trazione, poteva capitare che i rematori liberassero proprio in quel momento effluvi non troppo profumati.

Era necessario che i movimenti dei tre uomini seduti in scala fossero coordinati con grande precisione. Dalla vista laterale di una trireme si può vedere che i remi di ciascun terzetto di rematori devono mantenere durante la voga una posizione parallela, con uno scarto non superiore a 30 centimetri; in caso contrario si verificherebbero intralci che comprometterebbero il sincronismo della palata, coinvolgendo la maggior parte dei remi dello stesso fianco e alterando in tal modo pericolosamente non solo la velocità della nave, ma anche la sua direzione.

Polireme romana tipica, forse una quinquereme, illustrata in un mosaico del I secolo a.C. (particolare). Il mosaico è stato fotografato da Casson a Palazzo Barberini a Palestrina, vicino a Roma. I romani ruppero con la preferenza accordata dai greci a navi da guerra veloci. Ispirandosi ai cartaginesi, essi svilupparono infatti poliremi dallo scafo molto largo che potevano trasportare un gran numero di fanti. Rispetto alla trireme ateniese classica, che trasportava solo 14 fanti, le quinqueremi romane ne contenevano sino a 120. Nelle navi costruite in epoca posteriore pare che il posticcio sia stato ingrandito e anche rafforzato in modo da poter sostenere tutti i remi e non solo quelli del banco superiore.

Per controllare in modo appropriato il sistema, era necessario un addestramento continuo, e una grande importanza veniva attribuita all'incentivazione. Nell'antichità i vogatori venivano ingaggiati principalmente fra i cittadini liberi, i quali erano interessati alla sopravvivenza della loro città-stato *(polis)*. In generale l'uso di schiavi era limitato a situazioni di emergenza, e anche allora essi venivano usati di solito solo dopo essere stati liberati. Ci si può rendere conto facilmente che la presenza di un solo individuo scontento su una nave poteva comprometterne l'efficienza in un momento critico. La sferza non fu mai usata; un capovoga dava a voce il tempo della voga. Un altro incentivo era costituito dal compenso relativamente elevato che i rematori ricevevano in cambio delle loro prestazioni. L'efficacia di quest'incentivo era tanto maggiore in quanto i vogatori venivano ingaggiati di solito nelle classi più povere delle città, per le quali il costo delle complesse armature indossate dalla fanteria pesante era inaccessibile.

Si può immaginare che fra i vogatori che faticavano sui banchi e i soldati dall'armamento pesante che stavano inattivi sul ponte in attesa di combattere esistesse una certa tensione sociale. Solo negli stati più democratici, come Atene, si contò soprattutto sull'azione distruttiva dello sperone e si ridusse a un minimo il contingente di uomini armati. Nelle triremi impiegate nella guerra del Peloponneso gli ateniesi trasportavano solo 14 uomini armati. La sistemazione compatta dei rematori permise, per contrasto, di comprimere 170 vogatori in uno scafo che non era molto diverso in lunghezza e in larghezza da quello della pentecòntoro.

Il risultato di questo sforzo complesso va visto in un livello sorprendente di velocità e di manovrabilità. Stime ragionevoli della velocità massima di una trireme parlano di 11,5 nodi (21,3 chilometri all'ora). In effetti queste stime potrebbero essere anche troppo prudenti, poiché noi supponiamo che lo scafo avanzando spostasse acqua, mentre taluni ingegneri navali sostengono che le triremi erano abbastanza leggere e veloci da trarre vantaggio da una condizione di quasi scivolamento, come quella che si riscontra nelle imbarcazioni veloci di oggi. In tali condizioni la velocità potrebbe essere stata superiore del 50 per cento. Tali velocità potevano essere mantenute solo per cinque o dieci minuti prima che i vogatori si stancassero, ma in tali brevi intervalli di tempo potevano reggere il confronto con le velocità ottenute nelle cariche dalla cavalleria pesante del Medioevo.

Le triremi impiegavano probabilmente una trentina di secondi per raggiungere la velocità massima dopo una partenza da fermo, ma poiché una velocità intermedia poteva essere raggiunta in forse otto secondi e un quarto di tale velocità in due secondi, non ci voleva molto tempo per impartire alla nave un'energia sufficiente a infliggere un colpo decisivo a una nave nemica. Le triremi erano chiaramente navi molto efficaci nello speronamento. Inoltre potevano vantare una manovrabilità eccellente. Remando all'indietro su un fianco e in avanti sull'altro, per esempio, il loro equipaggio poteva farle ruotare in uno spazio di poco maggiore alla loro stessa lunghezza. Doveva essere davvero impressionante vedere una dozzina di queste navi sfrecciare nell'acqua nel corso di manovre o mentre gareggiavano, dopo un'esercitazione, per vedere quale di esse sarebbe entrata in porto per prima.

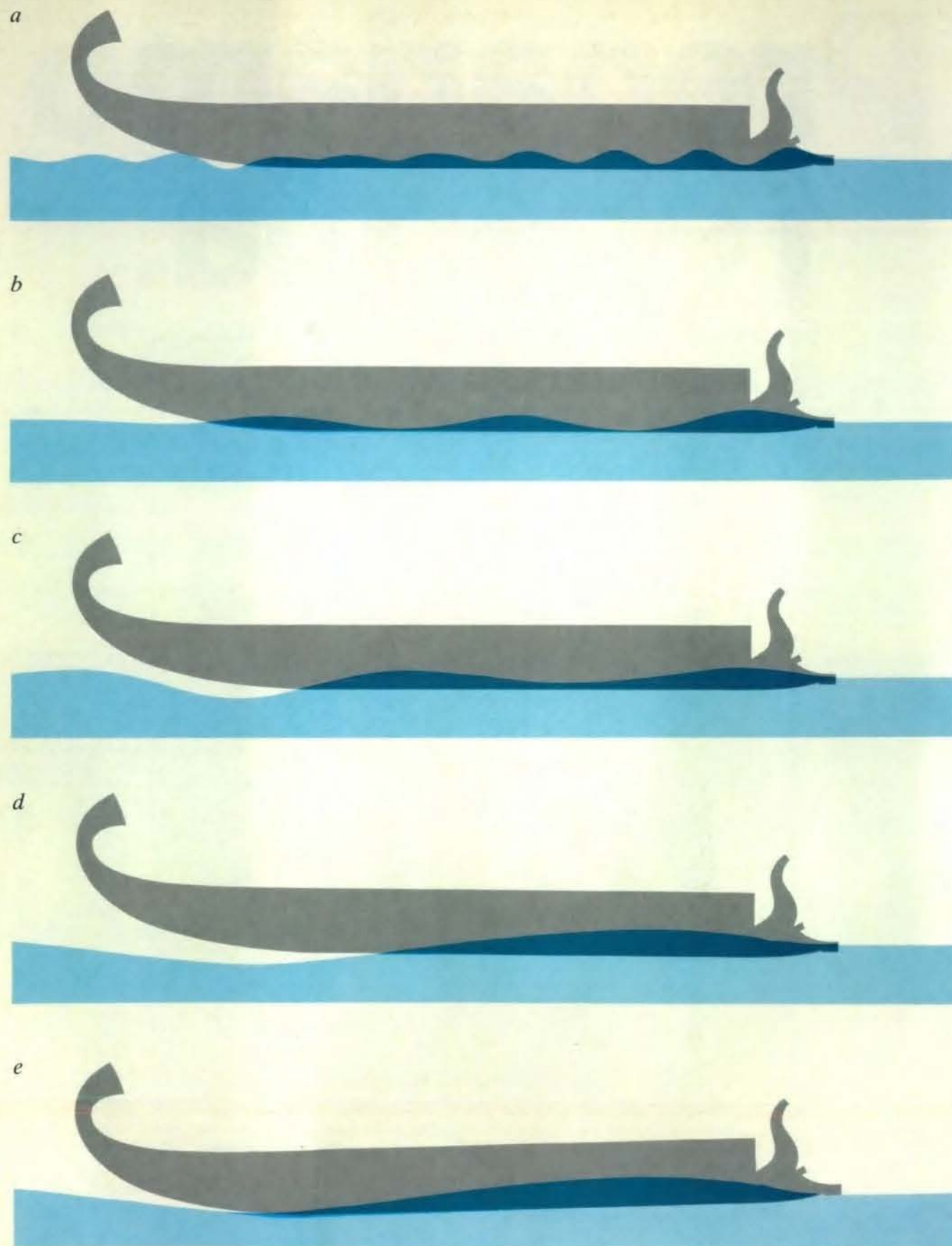
Una fra le prestazioni più famose su lunga distanza di una trireme del periodo di cui ci stiamo occupando ebbe luogo nel 427 a.C. La città di Mitilene, nell'isola di Lesbo, si era ribellata contro gli ateniesi, ma era stata poi riconquistata. Il demagogo ateniese Cleone propose allora che l'intera popolazione della città fosse messa a morte e la sua retorica si

a

b

c

d

e

Il sistema di onde creato dal passaggio di una nave nell'acqua è la forma più importante di resistenza dell'acqua al movimento della nave. A basse velocità *(a)* la formazione di onde incide molto poco sul movimento della nave. Al crescere della velocità aumentano sia la lunghezza d'onda sia l'ampiezza del sistema di onde. L'onda di prua può accentuare o attenuare l'onda di poppa, a seconda che i due sistemi siano fuori fase *(b)* o in fase *(c)*. Infine la cresta dell'onda di prua tenderà a sollevare la prua della nave nello stesso tempo in cui la depressione dell'onda di poppa fa abbassare la poppa della nave *(d)*. A questo punto la nave perderà il giusto assetto e si richiederà una frazione crescente della sua potenza per consentirle di risalire l'onda *(e)*. Per evitare il più possibile questa condizione, la nave va costruita il più possibile lunga.

impose nell'assemblea pubblica. Si affidò allora a una trireme l'incarico di andare a comunicare l'ordine alla guarnigione ateniese. Dato l'alto livello delle passioni politiche ad Atene, la trireme salpò probabilmente presto nel pomeriggio, subito dopo il voto dell'assemblea. Come scrive Tucidide, però, a causa della natura raccapricciante del messaggio che doveva portare, la nave navigò senza affrettarsi particolarmente. Impegnando uno o due ordini di rematori, e procedendo a un ritmo di voga lento, essa non percorreva probabilmente più di quattro o cinque nodi all'ora. La mattina seguente l'assemblea tornò a riunirsi; questa volta ebbero la meglio giudizi più spassionati e l'ordine di massacrare la popolazione fu annullato. Sperando in questo ripensamento, gli ambasciatori di Mileto ad Atene avevano già predisposto una nave veloce e un equipaggio formato da vogatori di prim'ordine, cui fornirono cibi molto energetici per il viaggio e cui promisero un forte compenso in denaro nel caso che fossero riusciti a raggiungere la trireme prima che l'ordine fosse stato eseguito.

La seconda nave partì per Lesbo, a circa 345 chilometri da Atene, circa ventiquattr'ore dopo la prima. Raggiunse il mare aperto prima di notte e i rematori continuarono a vogare con energia per tutta la notte. Essi venivano addirittura nutriti mentre remavano, con focacce d'orzo inumidite nel vino. La notte era abbastanza chiara per consentire la navigazione e non c'era vento di prua. Per poter mantenere massima la velocità, i comandanti o imbarcarono un numero di rematori sufficiente a dare periodicamente il cambio a uno dei tre banchi, oppure mantennero continuamente ai remi due banchi di rematori al completo concedendo periodicametne un turno di sonno al terzo. Ma qualunque sia stata la tecnica adottata, la nave giunse a Mitilene a mezzogiorno, subito dopo l'arrivo della prima nave. Pare quindi che questa seconda nave abbia impiegato meno di ventiquattr'ore a compiere il viaggio, procedendo a una velocità vicina a nove nodi (16,6 chilometri all'ora). Il messaggio era già stato consegnato, ma la guarnigione non aveva ancora avuto il tempo di eseguire l'ordine. Un traghetto moderno compie il viaggio in 14 ore.

Il successivo passo avanti nella progettazione di navi a remi fu compiuto attorno al 400 a.C. A quest'epoca furono sviluppate navi a quattro banchi di remi, chiamate quadriremi, e nel 399 un gruppo di ingegneri riuniti da Dionisio di Siracusa costruì la prima quinquereme, o nave a cinque banchi di remi. A tutta prima si potrebbe pensare che quadriremi e quinqueremi rappresentassero semplicemente la continuazione di una tendenza iniziata dalle navi a un solo ordine di remi e proseguita poi dalle biremi e triremi, ma non è così. Un certo numero di considerazioni indicano che nessuna nave a remi dell'antichità fu mai costruita con più di tre ordini di remi sopprapposti. Fisicamente questa conclusione è giustificata dal fatto che l'angolo formato dei remi con l'acqua, reso necessario dalla creazione di un quarto ordine sovrapposto di rematori, avrebbe reso molto gravoso lo sforzo di questi ultimi. Già su una trireme i rematori dell'ordine superiore faticavano assai più degli altri e ricevevano perciò una paga superiore.

Pare invece che questo sviluppo sia consistito nell'aggiungere un vogatore a uno o più banchi, cosicché una parte dei remi venivano manovrati ora non più da uno, ma da due uomini. Dato lo spazio minimo disponibile nello scafo della trireme, quest'innovazione fu adottata in principio, con ogni probabilità, per il banco superiore, trasformando così la nave in una quadrireme. Solo al livello più alto era infatti possibile trovare spazio sufficiente per alloggiare altre due file di rematori senza modificare in misura considerevole la struttura della nave. Un'impostazione alternativa pare sia stata adottata invece dai cartaginesi, le cui navi remiere seguivano probabilmente la preferenza dei fenici per scafi massicci; pare che i cartaginesi abbiano adottato sin dal principio uno scafo abbastanza largo da poter alloggiare quattro rematori fianco a fianco a ciascun livello. Pare però che la linea principale di sviluppo della nave a remi sia passata in principio per il tipo dallo scafo più stretto.

Il passaggio dalla quadrireme alla quinquereme richiese evidentemente una nuova progettazione dello scafo, poiché in caso contrario si dovrebbe supporre che i costruttori di navi si accontentassero di collocare i vogatori aggiunti in una posizione così vicina al fulcro del remo che i loro sforzi avrebbero potuto aggiungere ben poco alla velocità della nave. Il mutamento di progettazione più probabile è che i costruttori abbiano trasformato il posticcio per i remi fuori delle murate dandogli una maggiore profondità, così che esso potesse consentire il raddoppiamento del numero dei rematori non solo dell'ordine superiore, ma anche di quello di mezzo, fornendo l'appoggio a due ordini di remi. Il banco inferiore rimase probabilmente con un solo rematore per remo per preservare il più possibile i van-

Il moto della nave, che può essere distinto in guinata, beccheggio e rollio *(in alto)*, ha un forte effetto sulla direzione ottimale di tiro di un'arma come la catapulta. Nel caso delle antiche navi a remi, per esempio, l'alto rapporto tra lunghezza e larghezza le rendeva 50 volte più sensibili al rollio che al beccheggio. Poiché una catapulta, per poter colpire un bersaglio che si trovi a una distanza considerevole, dev'essere inclinata a un forte alzo, la punta della freccia di una catapulta sarà più lontana dal centro di rotazione della nave di quanto non sia la cocca *(in basso a sinistra)*. Al rollare della nave, la punta del proiettile si troverà perciò più spostata di lato rispetto alla cocca *(in basso a destra)*. Il puntamento iniziale poteva aver luogo facilmente solo quando la freccia e il bersaglio si trovavano su un piano verticale comune. Un proiettile scagliato quando la nave si sia discostata anche di poco dalla verticale andava a cadere lontano dal bersaglio. Perciò l'efficacia del tiro di una catapulta si riduceva di molto quando si tirava a un bersaglio nella direzione del moto della nave, per esempio a una nave che si tentava anche di speronare.

L'efficacia del tiro con catapulta da una nave che stava cercando di evitare lo speronamento veniva accresciuta dalle sue caratteristiche di stabilità e dalla forma del suo bersaglio. Poiché le lunghe poliremi erano particolarmente sensibili al rollio, i proiettili della nave sulla difensiva, tirati in direzione più o meno perpendicolare all'asse longitudinale della nave, subivano una dispersione considerevole in elevazione; tale dispersione dava origine a un'area di caduta dei proiettili in forma di un'ellisse allungata *(in alto e al centro)*. Poiché in questo caso la nave che stava attaccando era però a sua volta lunga e stretta, si aveva una buona corrispondenza fra la figura dell'area di caduta dei proiettili e la forma del bersaglio e quindi una maggiore probabilità di colpire quest'ultimo. (La disposizione a scalinata dei vogatori all'interno dello scafo accresceva a sua volta le probabilità che più di un uomo venisse colpito da un singolo proiettile.) Per contrasto, il rollio della nave che stava attaccando doveva avere come conseguenza una considerevole dispersione laterale dei suoi proiettili, dal momento che essa tirava più o meno nella direzione del suo moto *(in basso)*. Se la nave bersaglio offriva il fianco, questa dispersione non avrebbe dovuto avere conseguenze molto gravi, ma la piccola profondità da essa offerta al tiro faceva sì che anche piccole variazioni di alzo, come quelle causate dal carattere discontinuo della propulsione a remi, potevano rendere la traiettoria del proiettile troppo lunga o troppo corta. Il tiro effettuato con catapulte fra navi era a quanto pare considerevolmente più favorevole alla nave che si difendeva che non alla nave che attaccava.

taggi per la velocità connessi a una sezione immersa stretta, anche se ovviamente il pescaggio di una quadrireme o di una quinquereme non poteva non essere accresciuto dal peso dei rematori in più. Poiché il banco superiore nelle triremi alloggiava in origine 31 rematori per lato e gli altri due banchi 27 rematori per lato, le nuove quadriremi avevano probabilmente un totale di 232 rematori e le quinqueremi di 286.

I documenti iconografici più antichi di questa serie di mutamenti cominciano solo un secolo dopo l'introduzione di quadriremi e quinqueremi. In conformità con le richieste fisiche della voga, queste immagini confermano che le poliremi greche, di dimensioni maggiori, del III e II secolo a.C. avevano la maggior parte o la totalità dei loro remi appoggiati non sul bordo della nave, ma su un grande telaio che sporgeva a sbalzo dalle murate. I romani presentarono la tendenza a seguire la preferenza cartaginese per uno scafo più largo con semplici portelli per i remi praticati nelle fiancate. Essi non erano naviganti particolarmente abili e preferivano, ogni volta che ne avevano la possibilità, agganciare le navi nemiche, abbordarle e trasformare quindi le battaglie navali in scontri in cui si potessero impiegare le tattiche della fanteria, in cui essi eccellevano. Perciò per loro la velocità era meno importante della capacità di trasportare truppe.

Durante i primi decenni dell'esistenza di quadriremi e quinqueremi queste innovazioni si diffusero lentamente nelle marine da guerra dei paesi del Mediterraneo. Nel 330 a.C. Atene aveva nella sua flotta solo 18 quadriremi, contro 492 triremi. Nei sei anni successivi il numero delle quadriremi fu però più che raddoppiato ed ebbe inizio la costruzione di quinqueremi. Nel valutare le ragioni della

lenta adozione iniziale delle poliremi più pesanti, è utile confrontare dettagliatamente alcune fra le caratteristiche fisiche di triremi, quadriremi e quinqueremi.

È ovvio dal disegno della trireme che uno fra i problemi principali che i suoi progettisti dovevano affrontare era quello di evitare un peso eccessivo a prua. Le condizioni che stabiliscono se una nave è stabile e se sarà capace di raddrizzarsi quando si inclina vengono considerate oggi in relazione all'altezza metacentrica della nave stessa. Questo concetto non risale a prima del Settecento, anche se la maggior parte dei princìpi fisici e matematici in esso impliciti erano noti ad Archimede. In effetti ci sono indicazioni del fatto che le ricerche di Archimede sui galleggianti sarebbero state motivate dallo studio dei problemi della stabilità delle navi. Consideriamo il caso di una nave tipica vista in sezione trasversale *(si veda l'illustrazione a pagina 100)*. Il suo baricentro si troverà in un qualche punto all'interno dello scafo e, per ragioni di simmetria, su un qualche punto della linea mediana.

Possiamo immaginare che la nave si comporterebbe come se la sua massa fosse concentrata in quel punto e venisse attratta in basso, verso il centro della Terra, dalla gravità. Se la nave si inclina da un lato, il baricentro di quella piccola sezione dello scafo che è immersa si sposterà nella medesima direzione. La spinta idrostatica dell'acqua agirà attraverso il centro di gravità della sezione immersa, spingendola verticalmente verso l'alto. Finché il centro di massa della sezione immersa si muove abbastanza nella direzione dell'inclinazione così che la sua forza di galleggiamento orientata verso l'alto possa compensare l'attrazione verso il basso della forza di gravità trasferendola sul lato depresso dello scafo, l'effetto netto dell'azione simultanea delle due forze sarà quello di far raddrizzare la nave. L'altezza metacentrica della nave è la distanza fra il suo baricentro e il punto in cui la forza di galleggiamento diretta verso l'alto interseca la linea mediana della nave. Come si può vedere dall'illustrazione, quanto maggiore è quest'altezza, tanto più stabile sarà la nave.

Oggi sappiamo abbastanza sulle dimensioni delle triremi e sulla posizione del loro equipaggio e di altri pesi all'interno dello scafo per poter stimare la loro altezza metacentrica. I valori che si ottengono sono soggetti a variare a seconda degli assunti su fattori come il pescaggio della nave, la quantità di legname usata nella costruzione dello scafo e la disposizione dei vogatori al suo interno. Anche piccoli spostamenti di uno o più di questi fattori possono determinare mutamenti di cinque o sei chilometri all'ora nella forza del vento capace di rovesciare la nave. Finché però si mantengono costanti i parametri da un tipo di nave all'altro, il rapporto fra i valori metacentrici sarà corretto anche se nessuno dei due mumeri ha una grandezza assolutamente precisa. Perché quest'impostazione rimanga valida si deve quindi supporre che nel passaggio dalla trireme alla quadrireme o alla quinquereme lo scafo e il suo contenuto siano stati variati il meno possibile. Poiché l'ingegneria nautica ha manifestato storicamente un senso di estrema prudenza e conservazione nell'operare tali mutamenti, quest'assunto sembra del tutto plausibile.

Il principale imponderabile in quest'analisi è la quantità di zavorra trasportata dalla nave, poiché nelle antiche poliremi era presente nello scafo abbastanza spazio da imbarcare una zavorra sufficiente a modificare in misura considerevole le caratteristiche di stabilità della nave. Nella stima della zavorra abbiamo scelto di mantenere lo scafo il più possibile leggero, in conformità con gli altri tentativi ovvi da parte dei maestri d'ascia dell'antichità di produrre uno scafo in grado di dare una risposta massima agli sforzi dei vogatori. La nostra stima è di 13 000 chilogrammi, una massa sufficiente a impedire a venti inferiori ai 60 chilometri all'ora di rovesciare la nave.

Il risultato di queste stime è che le triremi dovevano avere un'altezza metacentrica di circa 0,4 metri, che è un valore appena sufficiente e non certo generoso. Ne consegue che la prudenza dimostrata dagli antichi nell'evitare la battaglia a meno che l'acqua non fosse del tutto tranquilla era ampiamente giustificata. Nel caso della quinquereme l'aggiunta di quattro file di rematori in più ai due ordini superiori della nave crea un dilemma che mette la stabilità della nave in concorrenza con la sua velocità. Nel caso di uno scafo con una sezione sommersa paragonabile a quella della trireme e con gli altri assunti uguali a quelli formulati sopra, l'altezza metacentrica della quinquereme risulta essere di solo 0,1 metri. Una tale nave non avrebbe potuto offrire una grande resistenza alla pressione delle onde e sarebbe stata rovesciata da venti di velocità superiore a 30 chilometri all'ora.

Questa condizione chiaramente insoddisfacente motivò fortemente i costruttori di navi a modificare il disegno degli scafi. Per ragioni cui abbiamo già accennato, l'unica scelta valida che si offriva loro era quella di accrescere la larghezza delle navi. Quest'impostazione avrebbe avuto però come inevitabile conseguenza un sacrificio parziale della velocità, che era stata il principale obiettivo che li aveva indotti inizialmente a progettare navi con un maggior numero di banchi di rematori. Nel caso di una quinquereme, se lo scafo fosse stato allargato sino a raggiungere una stabilità uguale a quella di una trireme, il suo guadagno di velocità rispetto a una trireme sarebbe stato solo del 14 per cento e la sua accelerazione sarebbe stata inferiore. L'equipaggio, nell'ipotesi che venissero aggiunti 62 rematori nel banco più alto e 54 in quello intermedio, sarebbe aumentato di quasi il 70 per cento. Se la stabilità della quinquereme fosse stata ridotta come abbiamo postulato sopra, il suo guadagno di velocità sarebbe stato del 29 per cento.

In considerazione della complessità di questi problemi, si comprende meglio perché siano passati due terzi di secolo prima che si siano cominciate a costruire quadriremi e quinqueremi in quantità apprezzabili. L'attaccamento degli ateniesi alla loro tradizione storica delle triremi può contribuire a spiegare in parte il ritardo. Le navi che avevano vinto la battaglia di Salamina non potevano essere messe da parte a cuor leggero. Quest'analisi non spiega però il netto aumento della popolarità delle quadriremi e quinqueremi a partire dal 330 a.C. circa e la straordinaria corsa agli armamenti che ne seguì fra le potenze navali dell'antichità.

Lo sviluppo di navi dotate presumibilmente di sei vogatori per ogni gruppo di remi era stato intrapreso in Macedonia e a Siracusa già attorno al 350 o 340 a.C., ma ben poco se ne seppe sin dopo la morte di Alessandro Magno, avvenuta nel 323 a.C. Attorno al 315 uno dei successori di Alessandro, Antigono, stava costruendo navi con sette vogatori per gruppo di remi. Nel 301 suo figlio Demetrio il Poliorcete aveva nella sua flotta navi che avevano sino a 13 vogatori per gruppo di remi. Prima di morire, Demetrio fece costruire una polireme che portava tale numero a 16. I suoi rivali costruirono navi che erano ritenute pari o migliori di queste, ma stranamente alcune di esse avevano un numero di vogatori per gruppo di remi minore. Lisimaco, per esempio, costruì una nave a otto che si dice si sia opposta con successo a quella di Demetrio con 16. Questa gara indica che si facevano molti tentativi per accrescere il numero dei vogatori. Alla metà del secolo furono costruite una nave di 20 vogatori per gruppo di remi e due a 30 e questo processo culminò con una nave gigantesca a 40 vogatori che fu varata forse quattro decenni dopo.

Su queste navi si sa ancora meno che sulle triremi, e anche su di esse ci sono state violente controversie. Recentemente, però, sono state proposte spiegazioni soddisfacenti della loro struttura, specialmente per opera di Lionel Casson dell'Università di New York. Secondo Casson il raddoppiamento del numero dei rematori sarebbe giunto al suo compimento con le navi a sei, ossia col raddoppiamento del numero dei vogatori dell'ordine più basso. A quel punto divenne inevitabile allargare lo scafo delle navi, sia che i singoli costruttori di navi o le tradizioni delle città si fossero o meno già mossi prima in tale direzione. Dai sei vogatori per ogni gruppo di remi in su si dovette escogitare qualche nuova disposizione. Assegnando a ogni remo un numero di uomini via via maggiore, sino a un massimo di otto circa, si poteva ottenere un numero corrispondentemente maggiore di rematori. (Quando tali soluzioni tornarono di moda nel Rinascimento, si vide che otto era il numero massimo di rematori ancora utile per ogni remo.)

Quando a manovrare ogni remo c'erano più di due uomini, i vogatori non potevano più remare stando seduti. La nave era fornita di panche, ma per cominciare

una vogata gli uomini dovevano alzarsi in piedi, fare un passo avanti e (se il remo era molto lungo) salire su una sorta di sgabello per immergere la pala del remo in acqua. Per completare il movimento del remo dovevano poi ridiscendere e gettarsi all'indietro con tutto il loro peso sulla panca. A giudicare dalla posteriore esperienza europea, una tale disposizione degli uomini permetteva di avvalersi di un numero relativamente piccolo di uomini esperti, e gli uomini più all'interno erano quelli cui era affidato il compito più delicato. I problemi della coordinazione dei vogatori assunsero minore importanza e il maggiore sforzo fisico, reso necessario da remi lunghi ora 17,5 metri, ebbe come conseguenza una diminuzione delle prestazioni delle navi anche oltre i limiti già suggeriti dall'allargamento dello scafo.

Una volta supposta una disposizione che prevedeva un numero multiplo di uomini per ogni remo, il numero dei banchi diventa ancora più incerto di quanto già non fosse in precedenza. Una nave con quattro vogatori per ogni gruppo di remi poteva essere ora una trireme con due rematori al banco superiore, oppure una bireme con due uomini per ogni banco oppure una nave a un banco solo con quattro uomini per ogni remo. Analogamente, la nave di Demetrio a sedici potrebbe avere avuto due banchi di otto uomini ciascuno, o tre banchi con gli uomini distribuiti in una qualche combinazione in grado di dare sedici, come sei uomini nei primi due ordini superiori e quattro uomini al livello più basso. Ma ci furono anche altre ragioni, oltre a quella della stabilità, a far cadere la scelta su navi dallo scafo più largo e meno profondo.

Al di là della nave di Demetrio a 16 ci sono indicazioni del fatto che il numero dei rematori cambiò di significato anche per un'altra ragione, come testimonia il fatto che tale nave fu affrontata con successo da una nave a otto costruita da Lisimaco. Casson ha addotto un certo numero di ragioni che inducono a credere che questa nave a otto e tutte le altre navi con più di 16 vogatori per gruppo di remi fossero costruite come catamarani, ossia con due scafi gemelli collegati da puntoni. Nella nave di Lisimaco i vogatori erano forse distribuiti su due coperte in ciascuno scafo, con quattro uomini a ogni remo. Questa sistemazione è sufficiente a spiegare tutte le classi maggiori di navi da guerra a remi, comprese quelle da 40, che avevano tre banchi in ciascuno scafo, con forse otto, sette e cinque vogatori rispettivamente in ordine discendente, dando per ciascun lato di uno scafo un totale di 20 vogatori. La maggiore larghezza e profondità dello scafo di tali navi rese possibile costruirle più lunghe delle triremi. Le navi da 40 erano lunghe 128 metri e potevano alloggiare 4000 vogatori.

La maggior parte delle testimonianze sulle prestazioni di tali poliremi gigantesche suggeriscono che esse dovevano essere stabili, ma lente. Pare ragionevole supporre, sulla scorta dei risultati ottenuti prima per le quinqueremi, che le loro prestazioni declinassero costantemente al salire del numero dei vogatori per ogni gruppo di remi al di sopra di cinque. Una rozza stima indica che il rapporto tra la loro potenza e il loro dislocamento non era superiore a un sesto rispetto al valore corrispondente per le triremi. Perché, dunque, se il periodo anteriore era stato dominato dalla trireme e dal suo sperone, l'età di Alessandro e dei suoi successori si volse a navi di velocità inferiore e di minore manovrabilità?

Una risposta parziale è che la tecnica dello speronamento perse di popolarità e fu sostituita dall'aggancio e dall'abbordaggio. I romani, in particolare, trasportavano sino a 120 fanti di marina sul ponte delle loro quinqueremi. Questo numero implica, come confermano del resto le immagini che si sono conservate, che le loro navi erano molto larghe e che avevano solo uno o due banchi di vogatori. Le navi remiere gigantesche dalla struttura a catamarano erano in grado di trasportare un numero di uomini considerevolmente maggiore. La nave a otto di Lisimaco trasportava in coperta 1200 fanti, e quella da 40 trasportò una volta 2850 fanti e 400 marinai di coperta. È chiaro che in una battaglia navale decisa dalla fanteria di marina una di queste piattaforme giganti avrebbe potuto prevalere su qualsiasi numero di triremi, ciascuna delle quali poteva trasportare un numero di soldati compreso fra 15 e 30. Già le dimensioni e la struttura di una tale nave indicano che sarebbe stato impossibile circondarla con un numero di triremi sufficiente a trasportare una forza paragonabile a quella che poteva trovare posto sul catamarano.

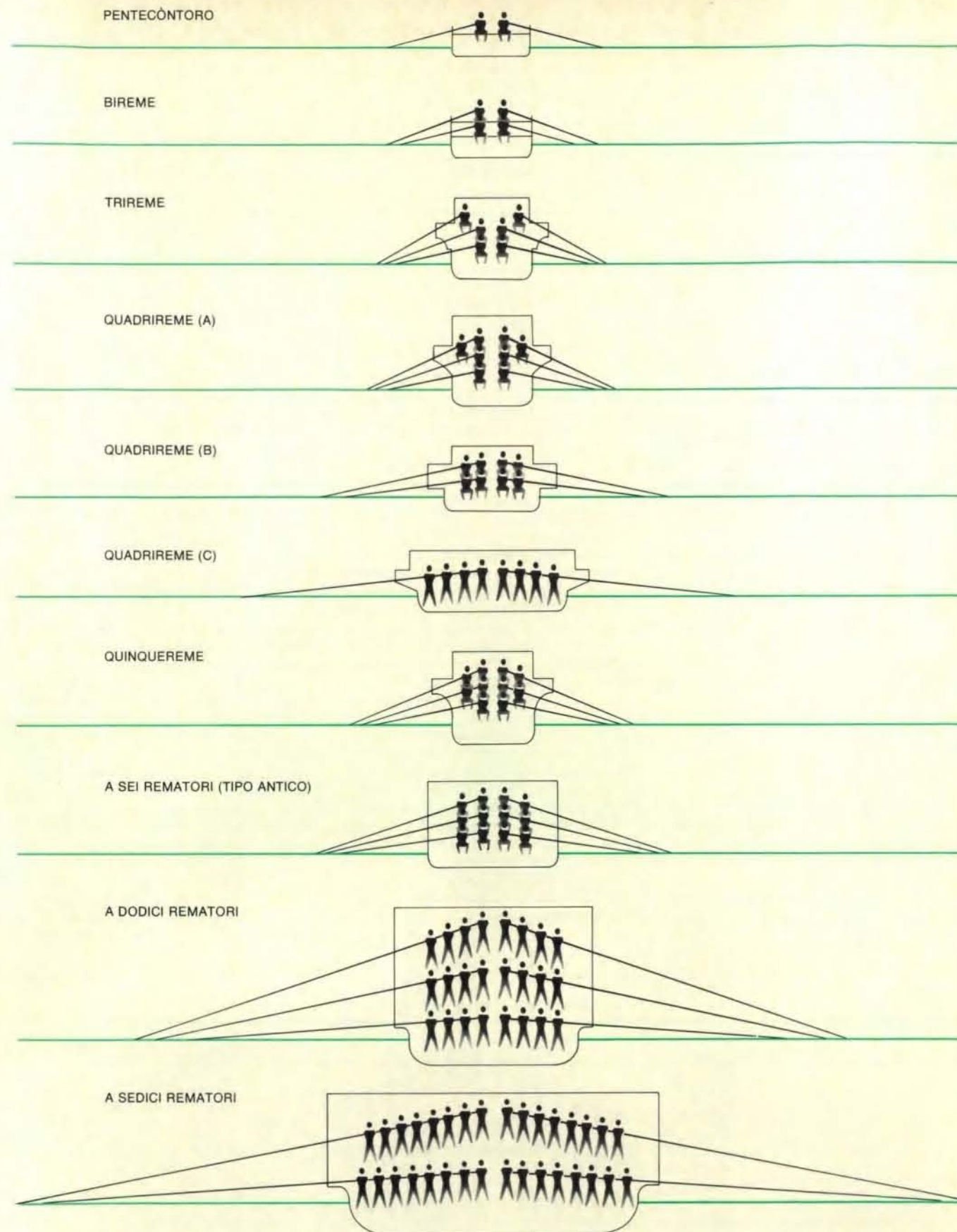
D'altra parte quest'argomentazione concede troppo poco spazio allo sperone. A meno che una tale nave avesse potuto essere circondata e difesa da una cortina di navi minori, i colpi infertile da una singola trireme potevano essere sufficienti ad affondarla. E se le sue capacità offensive erano limitate dalla velocità ridotta, perché tali navi venivano costruite e in dimensioni sempre maggiori? È chiaro che il passaggio a navi più lente e alle azioni di abbordaggio implica che attorno al 330 a.C. si era trovato il modo di neutralizzare lo sperone.

Pare che la risposta a quest'enigma debba essere vista in gran parte nelle antiche catapulte. Da molto tempo è stato rilevato che le navi più pesanti trasportavano catapulte e oggi sappiamo che la forma più potente e più sviluppata di quest'arma, che si fondava sulla forza elastica di corde fatte di tendini più che su archi flessibili convenzionali, fece la sua prima comparsa attorno al 332 a.C. Quell'anno, all'assedio di Tiro, Alessandro montò alcune delle sue catapulte pesanti su navi per tirare contro le mura della città. In seguito anche Demetrio, che portò il numero dei vogatori per ogni gruppo di remi, nelle navi remiere, da sette a 16, fece imbarcare catapulte.

Sottratte ai limiti che venivano imposti dal trasporto via terra, tali catapulte potevano raggiungere grandi dimensioni. Archimede costruì una catapulta per navi che scagliava pietre del peso di 78,5 chilogrammi o frecce lunghe 5,5 metri. Queste ultime venivano costruite evidentemente tagliando i rami ad alberi di medie dimensioni e armandone la cima con punte di ferro. La gittata della macchina con questi due tipi di proiettili era di poco inferiore ai 200 metri, ma catapulte minori potevano raggiungere, con proiettili più modesti, una distanza più che doppia. La maggior parte dell'attenzione dedicata all'effetto di tali proiettili sulle navi si è incentrata sulle palle di pietra. È stato espresso, con buone ragioni, il dubbio che questi proiettili potessero provocare l'affondamento di una nave. Anche se le palle avessero potuto sfondare le strutture superiori, lo strato di sabbia o pietre usato come zavorra avrebbe arrestato efficacemente il proiettile impedendogli di provocare danni gravi alla chiglia.

Quest'opinione trascura però la vulnerabilità dei fanti e dei vogatori ai proiettili delle catapulte. Per un tale attacco le frecce non erano meno efficaci delle pietre, tanto più da quando una nuova forma di punta di freccia era stata introdotta in Grecia attorno al 500 a.C. Una tale punta, di forma piramidale, aveva un grande potere di penetrazione. Giulio Cesare scrisse che le frecce scagliate dalle catapulte erano in grado di trapassare quasi un piede di legno di quercia. Le antiche poliremi offrivano certamente molto meno resistenza. Teofrasto riferisce che la coperta delle triremi era fatta di legno di tiglio, che ha qualità modeste di resistenza meccanica, ma è uno fra i legni più leggeri adatti alle costruzioni navali disponibili nel bacino del Mediterraneo.

Il nostro collega James F. Doyle, della School of Aeronautical and Astronautical Engineering alla Purdue University ha sperimentato gli effetti di una replica moderna di un'antica punta di freccia piramidale di dimensioni adatte per poter essere scagliata dalla più piccola fra le catapulte con corde di tendini la cui esistenza può essere documentata sulla base di antichi depositi di scarti di munizioni. Egli ha trovato che una tale freccia poteva penetrare per cinque centimetri nel legno di tiglio americano, la varietà americana del tiglio europeo, avendo un'energia equivalente a meno della metà della velocità che ci si poteva attendere che avesse quando fu lanciata. Questa stima è semmai troppo prudente, e la macchina usata era molto meno sviluppata della catapulta di Archimede già menzionata. Anche ammesso che il ponte delle navi fosse stato di quercia e avesse avuto lo spessore delle tavole dello scafo, le navi di dimensioni maggiori potevano trasportare catapulte multiple, in grado di scagliare frecce di grande potere di penetrazione, e potevano fare nondimeno ancora posto alla metà almeno del loro normale contingente di fanti di marina.

Non pare che le richieste di precisione che si ponevano al tiro con catapulte fossero eccessive. Nel Medioevo gli arcieri inglesi con archi normali praticavano il tiro al «*clout*», un pezzo di stoffa fissato

Lo sviluppo delle navi a remi nell'antichità passò per tre stadi principali. L'invenzione dello sperone, attorno all'800 a.C., condusse da semplici navi a un solo banco, come la pentecòntoro a 50 remi *(in alto)*, alla bireme e poi, con l'aggiunta di telai a sbalzo sporgenti dalle murate, i posticci, alla trireme. L'avvento della trireme mise fine in effetti alla prima fase di sviluppo, poiché non c'era praticamente alcuna possibilità di andare oltre tre livelli di vogatori. Il successivo passo avanti si concentrò sull'aumento del numero dei vogatori per ogni remo. Ebbe inizio una maggiore flessibilità di progettazione, in quanto i progettisti di navi avevano la possibilità di combinare in vario modo i numeri dei vogatori e dei banchi. Quest'impostazione permise probabilmente di giungere sino a un massimo di circa otto uomini per remo *(in basso)*, anche se in questo caso i vogatori dovevano stare in piedi e camminare avanti e indietro durante ogni vogata. Nella fase finale di questo processo evolutivo si ebbe la costruzione di catamarani giganteschi, o navi remiere a doppio scafo ***(si veda l'illustrazione alle pagine 110 e 111)*****.**

sul terreno, a distanze di oltre 200 metri. Non conosciamo le dimensioni precise di tale pezzo di stoffa, ma il suo diametro era certamente inferiore alla larghezza, di circa 5 metri, del ponte di una trireme. Rolfe Smith, il presidente dei National Crossbowmen, l'associazione dei balestrieri degli Stati Uniti, riferisce di essere in grado, con equipaggiamenti moderni, ma tirando con naturalezza, senza adottare particolari accorgimenti, di far cadere le sue frecce entro un raggio di tre metri a 365 metri di distanza. Poiché le antiche catapulte di grandi dimensioni erano montate su piedestalli regolabili, erano probabilmente in grado di avvicinarsi a tale precisione. Il puntamento doveva inoltre essere facilitato dalla visibilità delle frecce di grandi dimensioni durante il loro volo.

Una volta trapassata la coperta di una trireme, una freccia di catapulta aveva ottime probabilità di colpire un vogatore, data la compattezza con cui gli uomini erano sistemati ai remi. In effetti la disposizione a scalini dei vogatori consentiva a frecce sufficientemente grandi e lunghe di colpire più di un vogatore, tanto più che l'angolo di caduta delle frecce scagliate da lunga distanza era particolarmente favorevole allo scopo. Nei casi in cui c'era un solo uomo per remo, anche un solo colpo messo a segno poteva compromettere per qualche secondo il sincronismo della voga su un fianco della nave, tanto più se l'uomo colpito si trovava sul banco superiore, quello più vulnerabile, e immergeva il suo remo in mezzo agli altri. Anche nel caso che nessun vogatore fosse stato ferito, l'improvvisa presenza di una grande freccia nello spazio sovraffollato di vogatori avrebbe avuto molto probabilmente l'effetto di interrompere l'azione di qualche vogatore, con risultati in gran parte simili. Un ritardo di alcuni secondi poteva avere gravi conseguenze, dato che una trireme poteva percorrere uno spazio pari alla sua lunghezza in meno di sei secondi.

Pare perciò che ci siano stati motivi a sufficienza per modificare il disegno delle navi allo scopo di ridurre a valori minimi il rischio connesso al bombardamento da parte delle catapulte. La sostituzione di remi manovrati da singoli vogatori con remi più lunghi, azionati da più uomini, avrebbe ridotto il rischio della perdita di sincronismo nella voga nel caso che fosse stato ferito un uomo solo. La forza combinata di sei o otto uomini sarebbe probabilmente bastata a spezzare il fusto di una freccia grande quanto un giavellotto caduto in mezzo a loro. Un numero minore di banchi avrebbe presentato un bersaglio molto meno denso ai proiettili. Perciò, da questo punto di vista, le tendenze principali che si riscontrano nella progettazione delle navi a remi dopo il tempo di Alessandro appaiono giustificate.

Anche considerazioni connesse alla stabilità contribuiscono a spiegare l'abbandono della tecnica dello speronamento. Per un certo numero di ragioni esso era particolarmente efficace quando una nave poteva essere investita nella fiancata o nel quarto di poppa. Lo speronamento otteneva perciò i suoi massimi effetti distruttivi quando una nave, colta di sorpresa, subiva l'urto nella fiancata senza poter far nulla per evitarlo. Quando però le navi impiegavano catapulte, la situazione tattica poteva quasi rovesciarsi. La nave ferma lanciava bordate ad angolo retto rispetto al suo asse longitudinale. Poiché la nave aveva una stabilità modesta, il suo tiro veniva disturbato dal rollio, che causava variazioni nell'elevazione dei proiettili lanciati. Le variazioni nell'elevazione avevano a loro volta come conseguenza variazioni nella gittata. La grande lunghezza della galera ne minimizzava il bersaglio, che si sarebbe tradotto in deviazioni laterali nel suo tiro. Poiché la nave in avvicinamento era circa 10 volte più lunga che larga, presentava un bersaglio adattato quasi perfettamente alle caratteristiche di tiro della nave che si difendeva.

La nave che attaccava, d'altra parte, era costretta a tirare con le sue catapulte lungo la linea della sua direzione di movimento. Il suo tiro non veniva disturbato molto dal beccheggio, ma il bersaglio che essa cercava di colpire aveva una profondità di soli tre metri circa, esclusi i remi. Per la nave lanciata nel tentativo di speronamento costituiva invece un problema molto grave il rollio, perché quando si dava alle catapulte un alzo sufficiente per consentire ai proiettili di raggiungere un bersaglio lontano, la punta delle frecce veniva a essere più lontana, dal centro di rotazione della nave, dell'estremità con impennaggio.

Ne segue che, quando la nave rollava, la punta della freccia veniva deviata di lato più della cocca. Se il puntamento veniva fatto quando la nave era verticale, che era l'unico momento in cui si potesse puntare facilmente, e il tiro avveniva quando la nave aveva superato la verticale, le frecce sarebbero passate lontane dal bersaglio. Il moto di rollio è più veloce proprio alla metà del rollio stesso, cosicché determinare l'istante esatto in cui la nave passava per la verticale era estremamente difficile. Per una catapulta delle dimensioni considerate sopra, un errore di solo 1,5 gradi dalla verticale sarebbe stato sufficiente a mancare il bersaglio quando si tirava su una trireme in avvicinamento da una distanza di 200 metri.

Poiché abbiamo supposto che la nave bersaglio fosse immobile in attesa dell'attacco e che offrisse all'attaccante la fiancata, questo problema della deviazione del tiro potrebbe non apparire tanto grave. Anche il rollio della nave doveva incidere, per quanto in piccola misura, sull'elevazione della catapulta, così come la lieve sbandata della nave a ogni remata dei vogatori. Nonostante il grande rapporto tra lunghezza e larghezza della polireme, la prua si sollevava leggermente durante la palata, e si abbassava subito dopo. Anche col mare calmo queste irregolarità potevano essere determinanti quando si tirava a un bersaglio basso a lunga distanza. Inoltre, quando la nave bersaglio era immobile in acqua, eventuali perdite inflitte ai vogatori avrebbero avuto conseguenze meno disastrose di quelle sofferte da una nave in movimento.

Pare quindi, in complesso, che le catapulte fossero in grado di neutralizzare almeno parte della minaccia rappresentata dagli speroni e dalle tattiche di abbordaggio, tornate di moda. Le catapulte fecero la loro comparsa al tempo giusto per spiegare i mutamenti osservati nella disposizione dei vogatori; i loro livelli di prestazione erano sufficienti a mettere fuori combattimento i vogatori se non ad affondare le navi e le loro caratteristiche di fuoco da piattaforme su navi erano tali da scoraggiare ogni tentativo di speronamento. L'accresciuta stabilità dei catamarani li rese particolarmente idonei come piattaforme per catapulte. Rimane da spiegare perché, dopo il 250 a.C., tali grandi navi da guerra tornassero a sparire di nuovo dalla scena, per essere sostituite da triremi o da navi ancora più piccole.

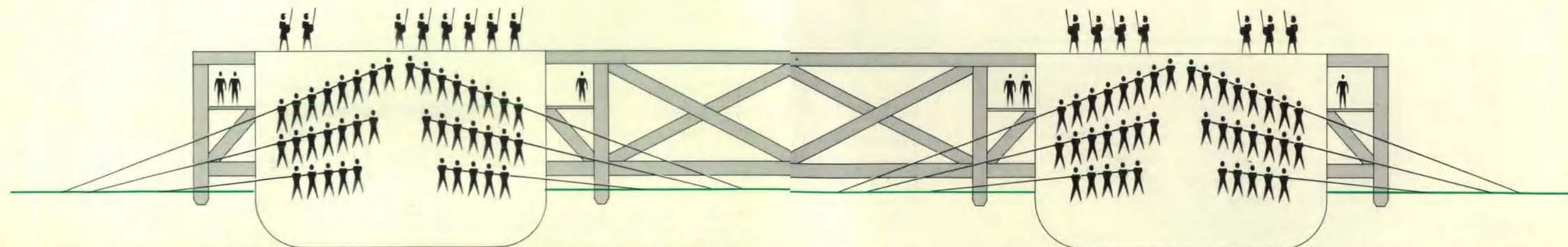
La ragione dev'essere stata almeno in parte tattica. Pare che la lentezza delle poliremi più grandi abbia incoraggiato l'evoluzione di tattiche di disturbo da parte di navi più piccole e più veloci. Alla battaglia di Chio, nel 201 a.C., per esempio, piccole navi aperte chiamate *lemboi* furono efficaci nel sottoporre a ripetuti attacchi le più pesanti navi di Rodi. Dopo l'attacco iniziale, quando le navi più grandi non si trovavano più nella formazione ordinata iniziale, i *lemboi* si insinuarono fra loro, ne danneggiarono o impacciarono i remi e ne spaccarono i timoni laterali. Un'insufficienza strutturale della catapulta le impediva di tirare a bersagli che si trovassero ad angoli inferiori all'orizzonte; ciò significa che una zona di sicurezza parziale circondava una grande nave armata di catapulte per quelle imbarcazioni che fossero state così piccole da insinuarvisi. Pare purtroppo che non esista però alcuna attestazione di piccole imbarcazioni che siano riuscite a insinuarsi fra gli scafi gemelli di catamarani, protette in parte dai posticci, distruggendone i remi su entrambe le fiancate.

Questa è un'idea interessante, ma sappiamo che almeno parte dei catamarani avevano gruppi di speroni che sporgevano fra i due scafi, impedendo l'ingresso nello spazio intermedio.

Il declino delle grandi navi a remi può essere ricondotto in parte anche ai grandi eventi politici del tempo. Con l'espansione di Roma, che mise fine alle rivalità fra i successori di Alessandro, la marina romana limita sempre più la sua azione a compiti che richiedevano navi più piccole e più veloci, come quello di liberare i mari dai pirati. Nell'ultima grande battaglia navale dell'antichità, quella di Azio, nel 31 a.C., le navi pesanti di Antonio furono sconfitte dalle navi più leggere e più veloci di Ottaviano dirette dall'ammiraglio Agrippa. Quest'ultimo aveva aggiunto vari ritocchi tattici che contribuirono al successo, come il lancio di grappini per mezzo di catapulte e di proiettili incendiari, alcuni dei quali pare venissero lanciati nello stesso modo. Le dimensioni e la complessità dei nuovi imperi richiesero una maggiore versatilità alla loro marina, cui furono affidati compiti come blocchi navali e il pattugliamento dei mari. I costi molto maggiori delle grandi poliremi, sia in uomini sia in materiali, contribuirono a determinarne il declino nel periodo imperiale. Anche Roma dovette ridurre al minimo le sue spese militari.

Una volta iniziato il declino di Roma, diminuì ancor più nella costruzione di navi l'insistenza su un uso sofisticato dell'energia muscolare umana. Dopo il 325 a.C. circa non si parla più di triremi. La nave da guerra a remi più diffusa dei periodi tardoromano e bizantino, il dromone, si affidava per l'offesa soprattutto al fuoco greco. Questo antico agente incendiario poteva essere proiettato per mezzo di una sorta di lanciafiamme oppure contenuto in un recipiente che veniva lanciato con una catapulta. Di solito i dromoni avevano due banchi, e ogni remo era azionato da uno, due o tre vogatori.

La successiva importante innovazione nelle navi a remi si ebbe all'inizio del Trecento, con lo sviluppo delle galee quando gli italiani introdussero il sistema di voga *a zenzile* (*sensile*, cioè singolo), che raggruppava i remi in gruppi di tre, ciascuno dei quali era manovrato da un solo uomo, come nella trireme. La novità consisteva nel fatto che i tre uomini erano seduti l'uno di fianco all'altro su una panca comune, collocata obliquamente nella nave in modo da poter vogare senza disturbarsi reciprocamente. Probabilmente la ragione principale per l'introduzione del nuovo sistema va vista nel centro di gravità più basso e nella maggiore stabilità che esso consentiva rispetto alle navi a due banchi. In Europa si era diffusa da poco tempo la bussola, la quale permetteva ora di navigare a vela con maggiore fiducia con tempo meno buono; le galee che scortavano i convogli dovevano essere in grado di avventurarsi in alto mare in condizioni che sarebbero state proibitive per le loro antenate greche.

Il mutamento finale nella progettazione delle navi a remi ebbe luogo attorno al 1550, quando il bisogno di trasportare pesanti cannoni costrinse a tornare alle galee dallo scafo lungo, con vari vogatori per ogni remo. Benché i cannoni trasportati su tali navi fossero montati in modo da sparare solo in avanti, il loro peso comportò molti mutamenti assimilabili a quelli associati, quasi duemila anni prima, al trasporto delle catapulte. Il futuro apparteneva però a navi da guerra in grado di sparare massicce bordate senza l'impaccio della presenza dei remi e dei loro vogatori lungo le murate. Dopo la battaglia di Lepanto, nel 1571, il ruolo delle galee declinò e dopo un'ultima fase di sopravvivenza limitata in mari interni come il Baltico esse scomparvero completamente. All'epoca della loro eclisse le navi a remi avevano alle loro spalle un'esistenza, nei mari del mondo occidentale civilizzato, durata quasi due millenni e mezzo: una sopravvivenza notevole dell'uso diretto della potenza muscolare umana in una cultura che si avvicina a essere dominata in misura sempre crescente dalla macchina.

Quattromila vogatori assicuravano la propulsione a questa mostruosa nave da guerra a due scafi fatta costruire da Tolomeo IV ad Alessandria verso la fine del III secolo a.C. Secondo la descrizione attribuita al greco Callisseno, conservata negli scritti di Ateneo e di Plutarco, la *tessarakònteres*, nave di quaranta rematori per ogni gruppo di remi, aveva due scafi ed era lunga 280 cubiti (128 metri); una volta trasportò 2850 fanti e 400 marinai di coperta. In questa ricostruzione degli autori, secondo uno schema elaborato da Lionel Casson dell'Università di New York, otto uomini sono assegnati a ciascun remo nei quattro banchi superiori, sette nei banchi di mezzo e cinque nei banchi inferiori. Accettando questa disposizione degli uomini in sezione, ogni fila longitudinale di uomini avrebbe dovuto comprendere cento uomini per dare il totale complessivo di 4000. Secondo i testi antichi, i remi più lunghi misuravano 38 cubiti (17,5 metri). La grande stabilità del catamarano ne fece la piattaforma ideale per grandi catapulte. Non c'è però alcuna prova del fatto che questa nave sia mai stata impegnata in battaglia.

Selezione sessuale nel bittacide dalle punte nere

Questa specie ha evoluto un singolare comportamento di accoppiamento: le femmine scelgono i maschi in base alle prede che questi presentano e i maschi possono imitare le femmine per rubare le prede ai rivali

di Randy Thornhill

Gli insetti, nonostante la relativa semplicità del loro sistema nervoso, presentano un'ampia gamma di comportamenti di accoppiamento differenziati e complessi. Poche attività di accoppiamento sono più singolari di quella del bittacide dalle punte nere, *Hylobittacus apicalis*, un insetto olometabolo dell'ordine dei mecotteri, con il corpo stretto e bruniccio, lunghe zampe esili e quattro ali strette e colorate di nero all'apice. Ho osservato per la prima volta il comportamento di questa specie una decina di anni fa, durante un'occasionale passeggiata in una zona boscosa del Michigan sud-orientale. Quando, all'inizio dell'estate, una popolazione di bittacidi adulti conta alcune migliaia di individui, non è difficile scorgerli mentre volano tra il sottobosco o si appendono alle piante con le esili zampe anteriori. I maschi usano le forti zampe posteriori per catturare afidi, mosche e opilioni, e io osservai che, dopo aver catturato una preda, ogni maschio ne staccava qualche piccolo boccone e poi, appeso a una foglia o a un rametto, attendeva che una femmina lo raggiungesse.

Quando una femmina si posava di fianco a un maschio questi le offriva la sua preda e spesso i due si accoppiavano; in alcuni casi, tuttavia, la femmina volava semplicemente via. Le femmine sembravano scegliere tra i maschi, rifiutando quelli che offrivano prede di piccole dimensioni e accettando quelli che ne offrivano di grandi. Mi venne in mente che questi insetti potessero fornire un eccellente sistema per mettere alla prova le idee di Charles Darwin sulla selezione sessuale, in particolare la sua concezione, frequentemente contestata, del ruolo della scelta da parte della femmina nell'evoluzione della morfologia e del comportamento del maschio.

Darwin sviluppò la sua teoria della selezione sessuale nel suo *The Descent of Man and Selection in Relation to Sex (L'origine dell'uomo e la selezione in rapporto al sesso)*, che venne pubblicato nel 1871. Egli dimostrava che l'evoluzione di caratteristiche sessuali secondarie può dipendere da due forze selettive distinte: la competizione tra i maschi per le femmine e la scelta da parte delle femmine di alcuni maschi a preferenza di altri. Nel primo tipo di selezione la *fitness* (o successo biologico) riproduttivo del maschio dipende dalla sua abilità nel vincere i conflitti con gli altri maschi della stessa specie, nel secondo caso dipende invece dalla sua abilità nell'indurre le femmine discriminanti a sceglierlo.

Secondo Darwin l'evoluzione di tratti maschili (come i complessi schemi di comportamento nel corteggiamento, l'elaborato piumaggio di certi uccelli, le corna di alcuni mammiferi ungulati e i processi simili a corna di certi insetti), che mettono in pericolo la sopravvivenza del maschio costandogli energia e aumentando la sua esposizione ai predatori, potrebbe venire spiegata in termini di vantaggi conferiti in uno di questi due processi selettivi. L'importanza della competizione tra i maschi come una forza nell'evoluzione dei tratti maschili non è mai stata messa seriamente in discussione, mentre l'importanza della scelta della femmina nella selezione sessuale è stata spesso discussa e messa in dubbio.

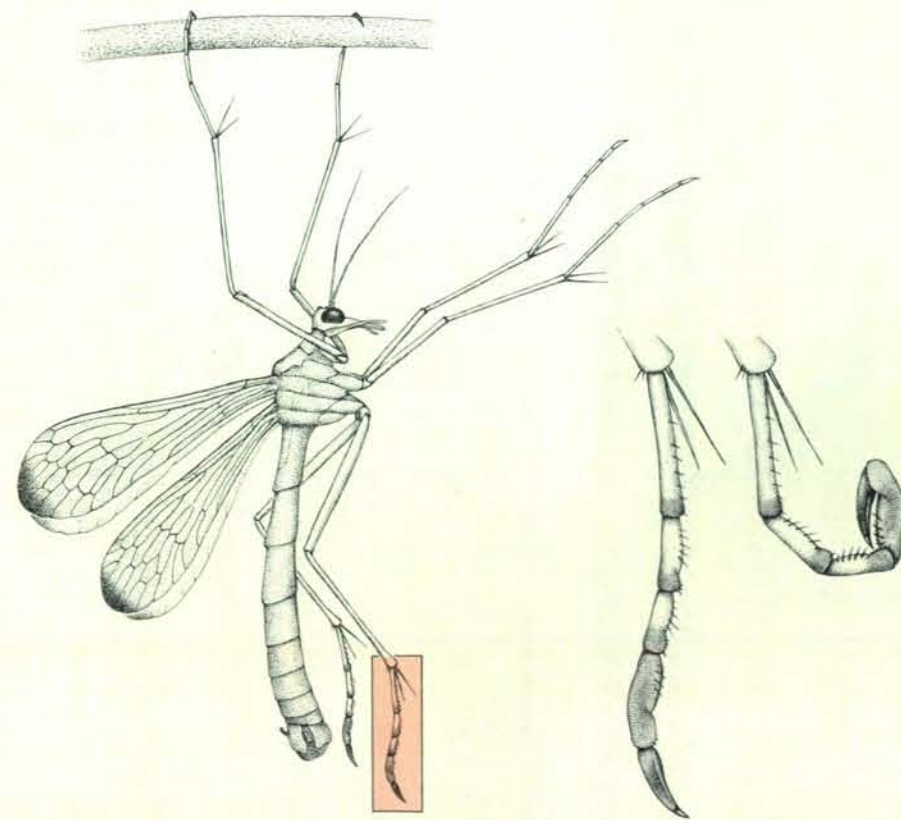
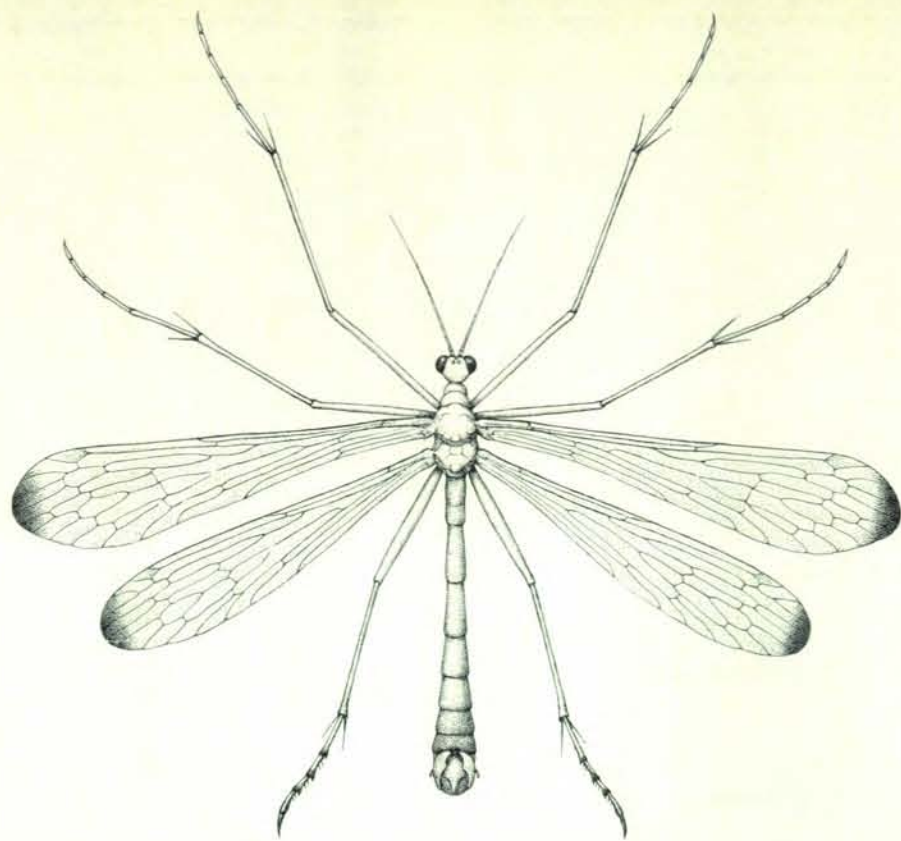
Alfred Russel Wallace, che giunse al concetto di evoluzione per selezione naturale indipendentemente da Darwin, fu uno dei primi a mettere in evidenza come non ci fossero abbastanza osservazioni in natura di scelte da parte della femmina per sostenere la teoria proposta da Darwin. Questa mancanza di evidenza diretta persiste ancora oggi: nella maggior parte dei casi l'esistenza di una scelta da parte della femmina viene dimostrata indirettamente per mezzo di studi che mostrano la distribuzione delle femmine intorno a maschi di diverse età o dimensioni oppure in possesso di differenti quantità di determinate risorse come il territorio. Oppure si suppone semplicemente che una scelta da parte della femmina abbia luogo in casi in cui le femmine si accoppiano con alcuni maschi e non con altri, nonostante il fatto che le femmine, essendo in genere più piccole, meno armate e meno aggressive dei maschi, possono spesso venire forzate a stare con certi maschi senza alcun riguardo per le loro eventuali preferenze. Inoltre ci sono poche prove che mostrino che la scelta da parte della femmina sia realmente adattativa in queste situazioni, servendo ad aumentare la sopravvivenza delle femmine che la praticano e, più specificamente, il numero di discendenti che esse generano.

A partire da quella mia passeggiata nei boschi del Michigan ho studiato il comportamento dei bittacidi e delle mosche scorpione, quest'ultime comuni anche in Italia, da un capo all'altro degli Stati Uniti e del Messico. Le mie osservazioni mostrano che le femmine della specie di bittacide *Hylobittacus apicalis* valutano veramente i maschi in base al dono presentato durante il corteggiamento: un insetto come preda. Non tutte le femmine di bittacide dalle punte nere scelgono tra i maschi in questo modo, ma quelle che lo fanno scelgono come compagni i maschi con le prede più grandi. Inoltre una tale scelta aumenta chiaramente le possibilità di sopravvivenza della femmina e il numero di uova che essa depone. Non vi è dubbio che il comportamento del bittacide dalle punte nere ci fornisce uno dei pochi esempi noti di scelta adattativa da parte della femmina. Per di più ho scoperto che i maschi di questa specie competono con le femmine con alcuni metodi degni di nota, compresa l'imitazione del comportamento femminile, per incrementare il proprio successo riproduttivo. Il bittacide dalle punte nere serve dunque come modello ideale per capire la selezione sessuale come Darwin la descrive.

Il bittacide dalle punte nere è anche un eccellente soggetto per uno studio sul comportamento in condizioni naturali. Lungo circa due centimetri, è attivo durante il giorno, catturando prede e accoppiandosi durante tutte le ore di luce. Gli individui possono venire marcati applicando una goccia di vernice sul corpo o sulle ali, e, nonostante i maschi e le femmine sembrino uguali, possono esser distinti dai differenti modi di volare: i maschi svolazzano avanti e indietro, mentre le femmine, appesantite da un ingente carico di uova, volano costantemente in linea retta. Gli individui di entrambi i sessi sono volatori relativamente lenti, effettuando solo brevi voli attraverso le piante erbacee che compongono il sottobosco del loro ambiente di foresta, e pertanto possono venire facilmente seguiti attraverso le loro frequenti interazioni con altri bittacidi e con gli insetti e gli altri artropodi di cui si nutrono.

Una sequenza di accoppiamento di *Hylobittacus apicalis* inizia quando un maschio cattura un artropodo oppure ne ruba uno a un altro maschio o a una coppia e, tenendo la preda con i tarsi posteriori, che in questa specie sono prensili, si aggrappa a una foglia o a un rametto e comincia a nutrirsene. (Per uccidere la preda il bittacide dalle punte nere fa uso della sua acuminata proboscide; quando viene conficcata nel corpo della preda, attraverso questa struttura simile a un becco, vengono iniettati degli enzimi che paralizzano la preda e ne liquefanno l'interno del corpo, che può poi venire succhiato.) Non tutte le prede catturate dai maschi vengono offerte alle femmine come dono nuziale. Le mie osservazioni di 42 maschi marcati in due siti di studio nelle foreste del sud-est del Michigan hanno rivelato che i maschi discriminano in modo evidente le prede che impiegano per questo scopo. Su 345 insetti che i maschi marcati avevano catturato, 110, cioè circa il 32 per cento, furono lasciati cadere sul tappeto di foglie che copriva il suolo della foresta dopo che i maschi se ne erano nutriti brevemente. Tutte le prede scartate misuravano meno di 16 millimetri quadrati di superficie corporea, il che suggerisce che fossero troppo piccole per rappresentare offerte adatte. (Nelle ricerche che sto qui discutendo, assumo come superficie corporea di un artropodo il prodotto della larghezza per la lunghezza; per esempio, una normale mosca adulta ha una superficie corporea di circa 20 millimetri quadrati.)

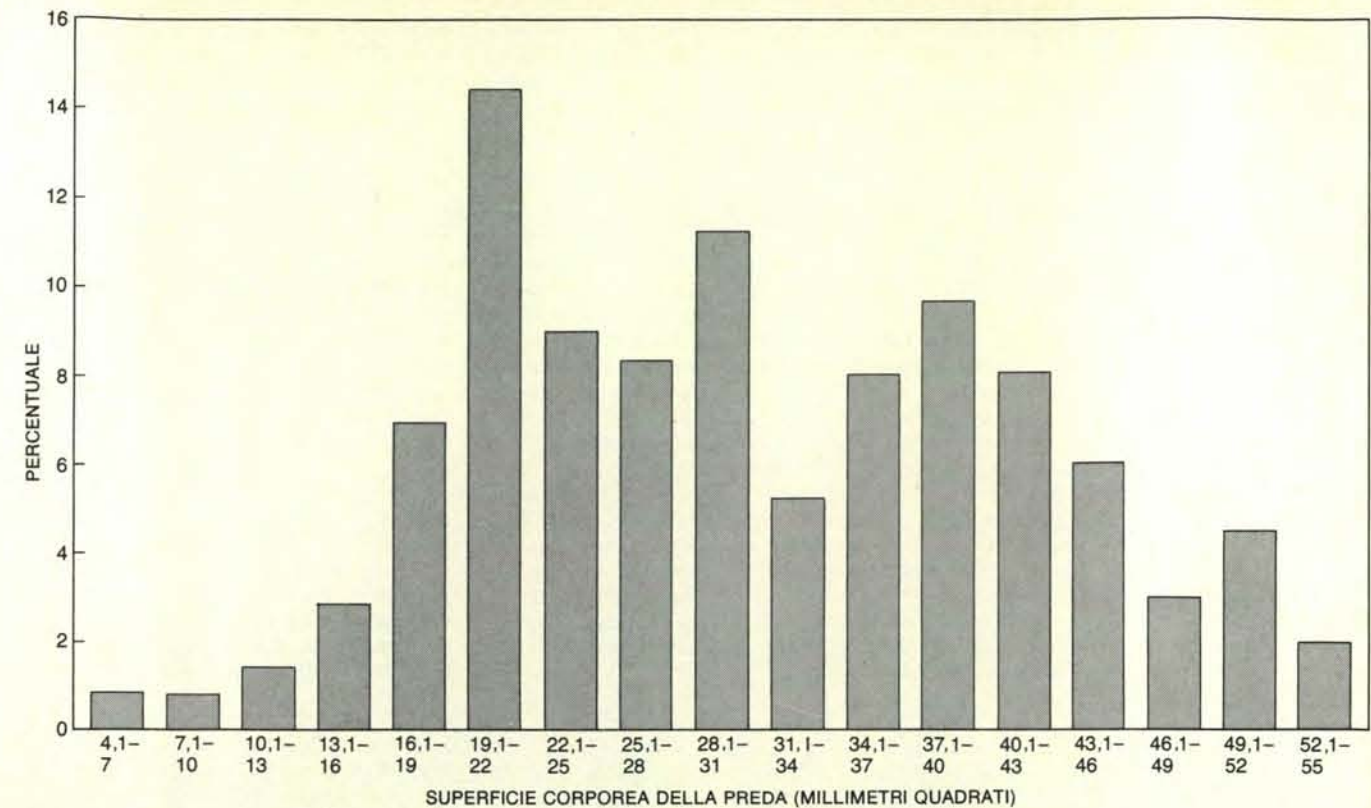
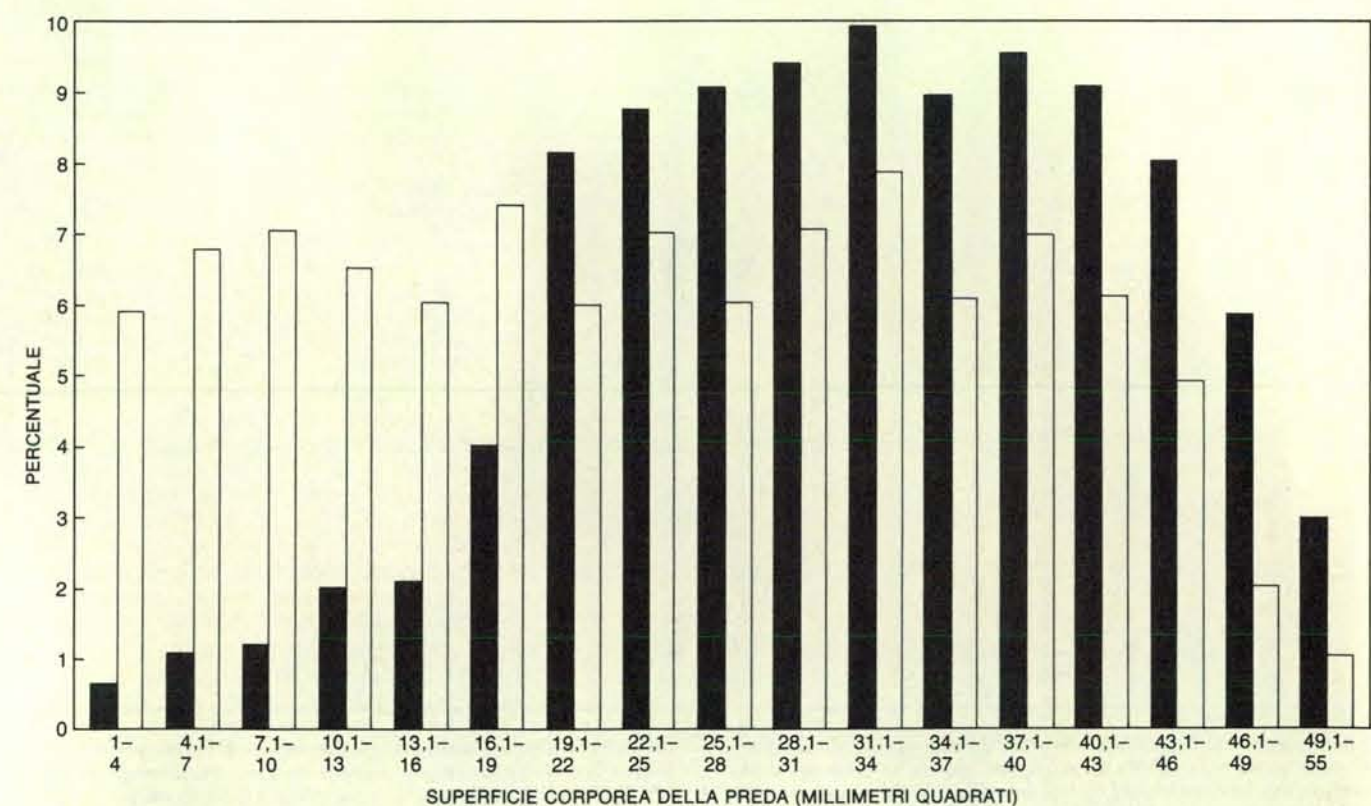
Un'ulteriore prova del fatto che i maschi trattengono la preda in base alle dimensioni fu ottenuta passando la vegetazione erbacea in ognuno dei miei siti di studio con un retino da insetti. Campionamenti di questo tipo furono effettuati ogni giorno per una settimana per determinare la distribuzione in base alle dimensioni degli artropodi disponibili. Inoltre, immediatamente dopo ogni campionamento furono catturate coppie in copula di bittacidi dalle punte nere e maschi in possesso di una preda, in modo che le prede potessero venire misurate. Il paragone fra questi campionamenti mostrò che nonostante i maschi di *Hylobittacus apicalis* disponessero di prede piccole (meno di 16 millimetri quadrati di superficie corporea) in proporzione diretta con la loro abbondanza nella vegetazione, erano in possesso di prede di grandi dimensioni (16 o più millimetri quadrati di

Questa fotografia mostra due bittacidi dalle punte nere *(Hylobittacus apicalis)* in accoppiamento mentre stringono il dono nuziale (un grosso moscone) che il maschio *(a sinistra)* ha offerto alla femmina *(a destra)*. Questi bittacidi sono nella prima fase del processo di accoppiamento, quando la femmina valuta la preda offerta dal maschio e il maschio tenta di accoppiarsi con lei. Una femmina di questa specie giudica i maschi in base alle dimensioni della preda offerta, rifiutandosi di copulare o copulando solo per breve tempo con i maschi che offrono prede piccole o di gusto sgradevole. Quando la preda è sufficientemente grande la femmina accetta il maschio e copula con lui per 20 minuti o più.

Il bittacide dalle punte nere, qui raffigurato dall'alto e di fianco, è un insetto di media taglia (circa due centimetri di lunghezza) che si trova nel sottobosco in aree ricoperte da foreste negli Stati Uniti orientali. Quando l'insetto non è in volo si appende con le zampe anteriori lunghe e sottili a una foglia o un rametto, come si può vedere in basso a sinistra. Per catturare e manipolare la preda il bittacide dalle punte nere fa uso dei forti tarsi prensili *(in colore)* delle zampe posteriori. Come si può vedere nell'ingrandimento in basso a destra, questi costituiscono un organo molto efficace per afferrare la preda, dal momento che si possono serrare per stringere un'ala o una zampa.

superficie corporea) e le offrivano come dono nuziale in proporzione molto maggiore rispetto alla loro abbondanza nell'ambiente naturale.

Nel mezzo dell'estate, quando i maschi di bittacidi dalle punte nere sono abbondanti, le femmine raramente catturano prede per proprio conto; esse dipendono quasi esclusivamente dai maschi e dalle prede che essi procurano. Le osservazioni fatte mostrano che quando le femmine cacciano, non scartano le prede di piccole dimensioni, e i miei studi confermano che al di fuori del contesto dell'accoppiamento, per quel che riguarda le femmine, la distribuzione di dimensioni delle prede coincide con la distribuzione naturale. Dunque sono solo i maschi ad essere predatori selettivi, catturando efficacemente prede di dimensioni ben precise, rispetto alla gamma di quelle disponibili.

Quando un maschio di bittacide ha catturato una preda sufficientemente grande, se ne nutre brevemente, poi, sempre tenendola con i forti tarsi posteriori, comincia a effettuare brevi voli attraverso la vegetazione erbacea. Nei maschi marcati ho osservato che questi voli prenuziali erano raramente più lunghi di 3,6 metri. Gli stimoli che inizialmente attraggono una femmina nelle vicinanze del maschio sono stimoli olfattivi, non visivi. Alla fine di ogni volo prenuziale il maschio si appende a una foglia o a un ramoscello, rovescia una coppia di sacchi ghiandolari dall'addome e inizia a liberare un ferormone, una sostanza che serve ad attirare le femmine. (Le ghiandole del ferormone vengono rovesciate per mezzo di un aumento della pressione sanguigna nell'addome del maschio e vengono retratte per mezzo di muscoli specializzati.) L'accoppiamento non comincia veramente fino al momento in cui la femmina non si appende, con le zampe anteriori di fronte al maschio. Allora il maschio presenta il dono nuziale alla femmina e, sempre mantenendo la presa, lascia che esso lo afferri con i tarsi posteriori. Non appena la femmina inizia a nutrirsi, il maschio tenta di accoppiarsi con lei.

Se il dono del maschio è grande abbastanza (16 millimetri quadrati di superficie corporea o più), la femmina lo accetta e i due copulano per una media di 23 minuti, durante i quali la femmina continua a nutrirsi. Il processo di accoppiamento è terminato dal maschio, e viene seguito da una lotta, quando il maschio tenta di sbarazzarsi della femmina e di liberare la preda dalla sua stretta. Se, al contrario, il dono del maschio è di gusto sgradevole (come avviene quando si tratta di una coccinella, che contiene sostanze disgustose), oppure troppo piccolo, la femmina rifiuta di accoppiarsi con il maschio o, se l'accoppiamento è già iniziato, lo termina rapidamente, volando via in media dopo cinque minuti. In quest'ultimo caso, poi, è la femmina che cerca di liberarsi dal maschio, ma se il maschio vince la lotta che segue per la preda, esso cerca generalmente di offrire nuovamente alla femmina il suo «indesiderabile»

I maschi preferiscono prede di grandi dimensioni da offrire come dono nuziale. L'istogramma mostra la distribuzione in base alle dimensioni degli artropodi usati da 168 maschi di bittacidi dalle punte nere nell'accoppiamento. Osservazioni di questa specie nel suo ambiente naturale indicano che più del 90 per cento degli accoppiamenti coinvolgono prede di grandi dimensioni, cioè insetti o altri artropodi che misurino 16 millimetri quadrati o più di superficie corporea. (Per gli scopi di questo studio la superficie corporea di un artropodo si ottiene moltiplicando la lunghezza per la larghezza; per esempio il moscone tenuto dai bittacidi a pagina 115 misura circa 30 millimetri quadrati di superficie corporea.)

Quando le femmine cacciano per proprio conto non prediligono prede di grandi dimensioni, come si può vedere in questo istogramma che paragona la distribuzione in base alle dimensioni delle prede catturate da 248 maschi *(in nero)* e da 276 femmine *(in bianco)* di bittacidi dalle punte nere. Le femmine sono generalmente ben rifornite di cibo dai maschi desiderosi di accoppiarsi e pertanto di solito non cacciano.

dono e di impegnarla ancora una volta nella copulazione.

I maschi sono stati visti vincere il 64 per cento delle lotte tra coppie per il possesso finale del dono nuziale e le femmine solo un 8 per cento. Nelle altre occasioni la preda era lasciata cadere e veniva persa nel tappeto di foglie sul suolo della foresta. Se il maschio conquista la preda dopo che la copulazione è terminata, se ne nutre brevemente e, se può ancora fornire adeguato nutrimento, la utilizza nuovamente per ripetere la sequenza di accoppiamento con un'altra femmina. In caso contrario la scarta e comincia a dare la caccia ad un'altra preda adatta come cibo nuziale. La possibilità che una preda possa venire utilizzata in due accoppiamenti successivi viene generalmente determinata dalle sue dimensioni. Ho trovato che gli artropodi che misurano da 28 a 45 millimetri quadrati di superficie corporea sono le prede che più frequentemente vengono utilizzate due volte. Prede più piccole non forniscono abbastanza nutrimento, mentre quelle più grandi per il maschio sono difficili da conservare, nella lotta che segue l'accoppiamento. In effetti sono le femmine ad avere più probabilità di mantenere il possesso di una preda di dimensioni molto grandi che è stata loro offerta come dono nuziale.

La femmina di bittacide dalle punte nere può dunque operare una discriminazione a sfavore dei maschi con piccole prede su due livelli: rifiutando tali maschi prima ancora che la copulazione inizi oppure accettando l'accoppiamento ma solo per un periodo limitato. I maschi con piccole prede in genere non riescono a copulare per più di cinque minuti, mentre quelli che ne offrono di grandi copulano per 20 minuti o più. La significatività di questi numeri non fu rivelata in modo chiaro fino a che non effettuai esperimenti in laboratorio per determinare quando, nel corso dell'accoppiamento, gli spermatozoi vengono veramente trasferiti dal maschio alla femmina. In questi esperimenti alcune femmine vergini furono fatte accoppiare per periodi compresi tra uno e 39 minuti. La spermateca (un organo dell'apparato femminile in cui viene immagazzinato lo sperma) venne poi estratta e fu applicato un colorante specifico per gli spermatozoi in modo che questi potessero venite contati.

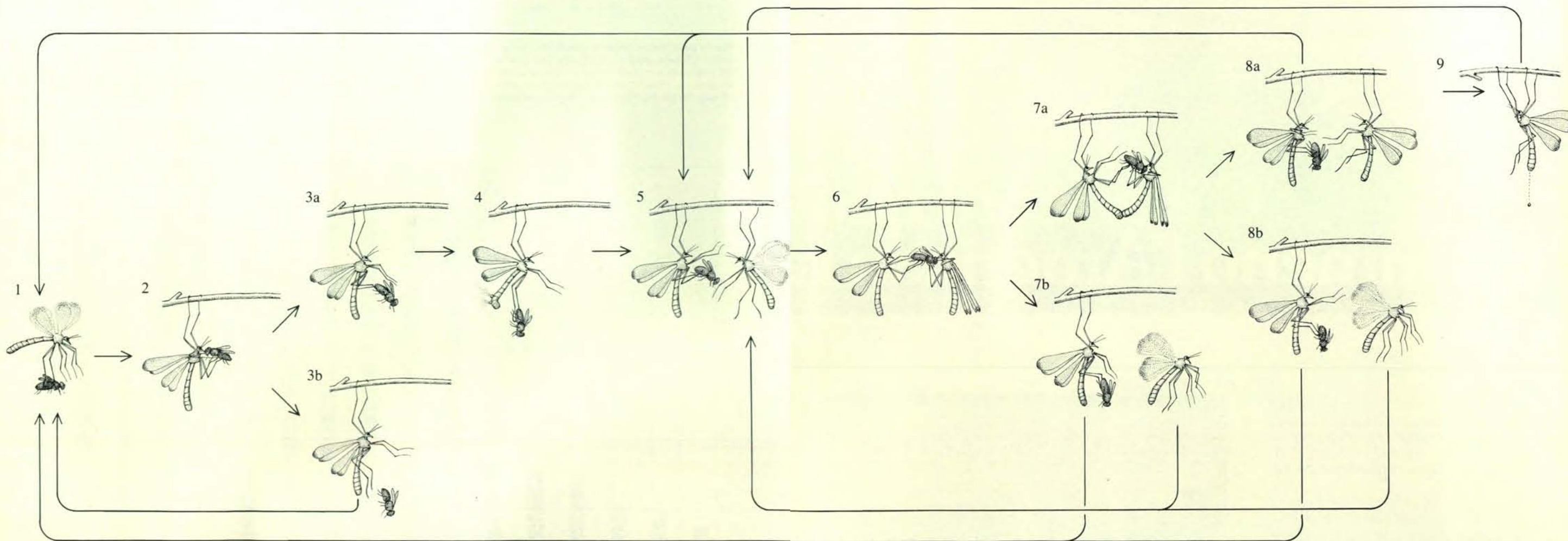
Questi studi hanno mostrato che sono necessari fino a 5 minuti di copulazione prima che gli spermatozoi inizino a passare nella spermateca della femmina. Per di più, per gli accoppiamenti che duravano da 5 a 20 minuti, ci fu una correlazione positiva tra la durata dell'accoppiamento e la quantità di spermatozoi trasferiti. Al di sopra dei 20 minuti c'era un piccolo aumento nella quantità di spermatozoi trasferiti, non importa quanto a lungo continuasse la copulazione. Quindi i brevi accoppiamenti dei maschi che portano prede troppo piccole hanno come risultato un trasferimento di pochi o addirittura nessuno spermatozoo, e dunque poco o nessun incremento numerico nella prole generata da questi maschi. I lunghi accoppiamenti dei maschi che portano grandi prede, al contrario, hanno come risultato la massima inseminazione della femmina, con positive conseguenze per la prole.

La massima inseminazione della femmina non è l'unico vantaggio che deriva al maschio con grandi prede dal suo periodo di accoppiamento più lungo. Osservazioni di bittacidi dalle punte nere effettuate sia in natura che in laboratorio hanno mostrato che il periodo di accoppiamento più lungo stimola la deposizione di uova e rende anche meno probabile che una femmina si accoppi nuovamente entro breve tempo. Le femmine di bittacide dalle punte nere depongono uova grandi e ben visibili, che restano attaccate all'addome alcuni minuti prima di cadere sul suolo della foresta. Con questo tipo di deposizione è possibile determinare la produzione di uova da parte delle femmine dopo interazioni con maschi che offrono prede di diverse dimensioni.

Ho osservato che dopo che una femmina ha copulato per 20 minuti o più essa non risponde ai maschi e non si accoppia nuovamente per circa quattro ore, durante le quali depone una media di tre uova. Se continua a incontrare maschi con grandi doni nuziali, questo ciclo di non ricettività all'accoppiamento e deposizione di uova può continuare ininterrottamente per tutto il giorno. Se, al contrario, una femmina si è accoppiata per solo circa cinque minuti, non depone nessun uovo ma trova rapidamente un altro maschio e poi un altro ancora, sempre in cerca di uno con una grossa preda. Non ha importanza la quantità di nutrimento che la femmina riceve dalle piccole prede nel corso di queste brevi interazioni; essa continua a interagire brevemente con diversi maschi fino a che non ne trova uno con una preda adatta e può quindi intraprendere un lungo accoppiamento.

La scelta da parte della femmina del bittacide dalle punte nere è chiaramente adattativa, cioè accresce la capacità riproduttiva delle femmine che la praticano. Ho osservato che le femmine che discriminano tra i maschi in base alle offerte nuziali depongono significativamente più uova di quelle che non discriminano. Inoltre le femmine che sono in grado di discriminare aumentano le proprie possibilità di sopravvivenza.

La ragione è la seguente. Un bittacide a caccia che si muove nel suo ambiente di foresta corre il rischio di volare in una ragnatela. Certamente una percentuale notevole di una popolazione di bittacidi viene perduta a causa della predazione da parte di ragni costruttori di ragnatele. Tra i bittacidi dalle punte nere, comunque, la percentuale di perdite è significativamente più bassa per le femmine, fenomeno che può essere attribuito alla disparità nelle distanze percorse dai due sessi nel corso delle loro normali attività. Tra gli individui marcati ho osservato che i maschi percorrevano una distanza media di circa 33 metri in un'ora, mentre le femmine nello stesso tempo si spostavano solo di circa 16 metri.

Questa disparità può esser fatta risalire alla pratica del cibo nuziale: i maschi passano circa il 50 per cento del loro tempo cacciando prede nuziali, mentre le femmine, essendo ben rifornite di cibo dai maschi, sono raramente costrette a cacciare.

Una sequenza di accoppiamento ha inizio quando un maschio di bittacide dalle punte nere cattura un artropodo oppure ne ruba uno a una coppia o a un altro maschio *(1)*, e poi, appeso a una foglia o a un rametto, se ne nutre brevemente *(2)*. Se la preda è troppo piccola, il maschio di solito la lascia cadere sul tappeto di foglie che ricopre il suolo della foresta *(3b)*, nel qual caso deve procurarsi una nuova preda e iniziare nuovamente. Se la preda è abbastanza grande, il maschio la tiene *(3a)* e, rovesciando una coppia di sacchi ghiandolari che si trovano nell'addome, inizia a liberare un feromone, sostanza che serve ad attirare le femmine *(4)*. Quando la femmina arriva, il maschio le offre la preda *(5)*, permettendole di afferrarla con i tarsi posteriori mentre anch'esso mantiene la presa *(6)*. La femmina inizia a nutrirsi della preda e la valuta mentre il maschio tenta di accoppiarsi. Se la preda è inaccettabile perché di gusto sgradevole o troppo piccola, la femmina la lascia andare e vola via, rifiutando di accoppiarsi con il maschio *(7b)*. Oppure la femmina può accoppiarsi con il maschio *(7a)*, ma solo per breve tempo, allontanandosi dopo circa cinque minuti *(8b)*. Se, al contrario, la preda è abbastanza grande, la coppia di bittacidi copula *(7a)* per una media di almeno 20 minuti, nel qual caso è il maschio a porre termine all'interazione *(8a)*. La femmina si nutre per tutto il periodo della copulazione, che termina invariabilmente con una lotta per il possesso della preda nuziale. (Il maschio in genere vince e, se la preda è sufficientemente grande, esso la utilizza una seconda volta come offerta nuziale per un'altra femmina.) Una femmina che ha copulato per soli cinque minuti trova rapidamente un altro maschio che le offre una preda e ricomincia il processo di accoppiamento. Una femmina che ha copulato per 20 minuti o più, invece, non risponde sessualmente ad altri maschi per circa quattro ore, durante le quali depone in media tre uova *(9)*.

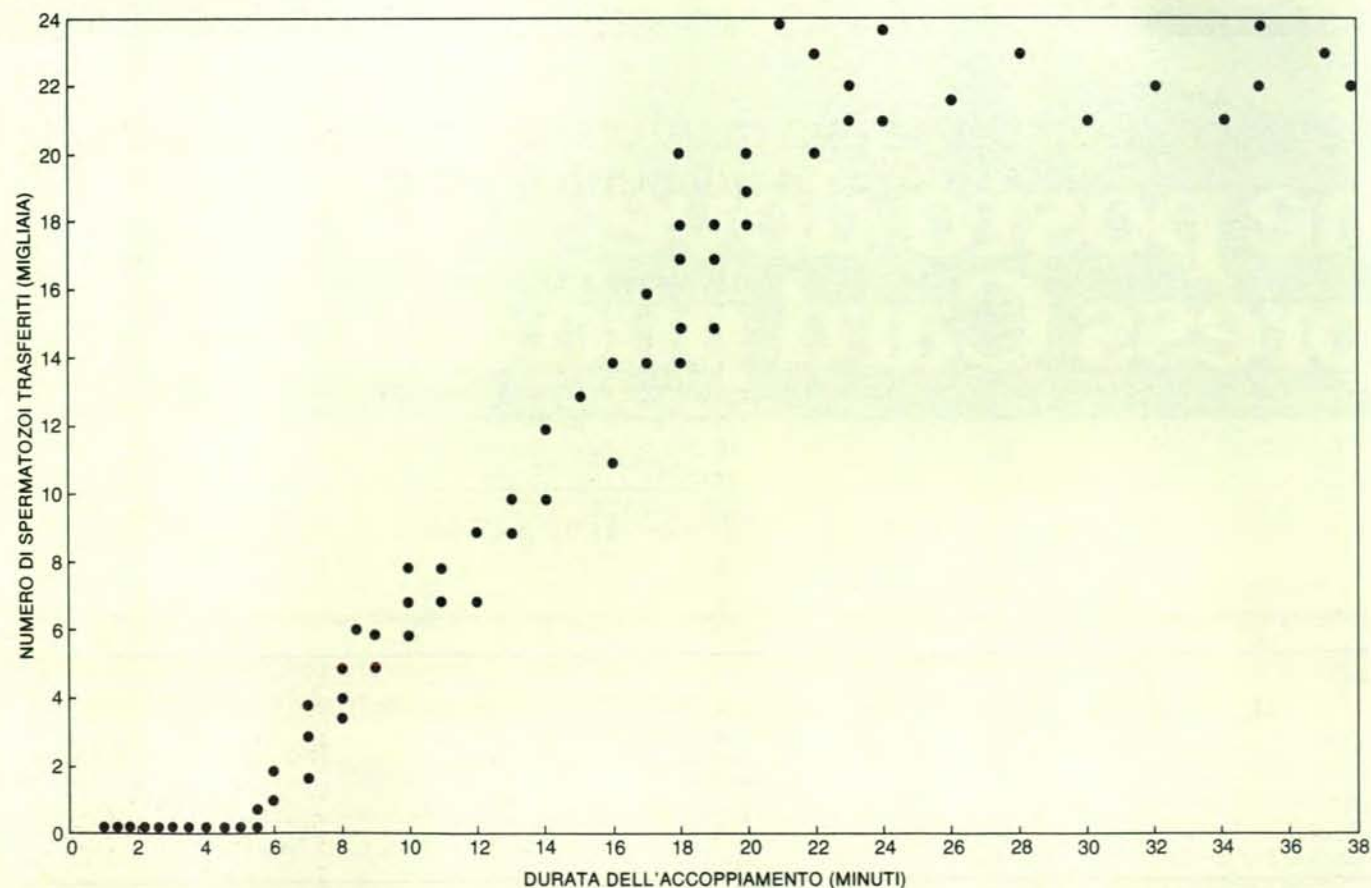
Perciò è nettamente meno probabile che le femmine incontrino delle ragnatele. Selezionando un maschio con una preda di grandi dimensioni, una femmina limita ulteriormente i propri movimenti, e quindi anche il rischio di essere catturata da un ragno costruttore di ragnatele, aumentando a causa di ciò le proprie possibilità di sopravvivenza. La relazione tra cibo nuziale e sopravvivenza della femmina del bittacide dalle punte nere è sostenuta dal paragone con un altro bittacide, *Bittacus strigosus*, che non usa il cibo nuziale. In questa specie maschi e femmine partecipano quasi equamente alla caccia e il tasso di mortalità a causa della predazione dei ragni nei due sessi è quasi identico.

Le mie osservazioni suggeriscono che il pasto nuziale si sia evoluto nel bittacide dalle punte nere come un mezzo per ridurre la partecipazione della femmina nelle attività che comportano dei rischi, vale a dire quelle associate con la caccia. Nel procurare cibo alla femmina il maschio riduce le probabilità che essa venga catturata da un ragno prima di deporre le uova che esso ha fecondato. Si deve notare che la mia interpretazione del vantaggio selettivo del pasto nuziale per i maschi è basata sulla supposizione, avvalorata da studi di altri insetti, che l'ultimo maschio che si accoppia con una femmina per 20 o più minuti feconda la maggior parte delle uova che essa depone.

Il fatto che la scelta da parte della femmina del bittacide dalle punte nere abbia come risultato che alcuni maschi (quelli con prede di grandi dimensioni) si accoppino più frequentemente e con maggior successo di altri (quelli con prede piccole) indica che la scelta da parte della femmina è potenzialmente una forma importante di selezione sessuale che forse esclude dalla riproduzione una frazione non indifferente di maschi in ogni generazione di bittacidi. Le mie osservazioni mostrano che in primo luogo, come risultato del numero limitato di prede disponibili, solo una piccola percentuale (tra il 2 e il 10 per cento) di maschi adulti in una popolazione di bittacidi è in possesso di una preda in uno stesso momento.

Di questi maschi si può giudicare che solo quelli osservati mentre effettuavano i caratteristici voli prenuziali o liberavano ferormone avessero valutato la preda allo scopo di accoppiarsi. In ogni momento il 90 per cento dei maschi che hanno valutato la preda sono in possesso di una preda di grandi dimensioni. In altre parole, il 10 per cento dei maschi che hanno valutato la preda porta prede piccole e, come risultato, viene scartato dalle femmine per l'inadeguatezza della preda. Dunque la scelta da parte della femmina può escludere dalla riproduzione fino al 10 per cento di tutti i maschi adulti di una popolazione. Inoltre la variazione nell'abilità dei maschi nel trovare rapidamente prede di grandi dimensioni impedisce a molti più maschi di raggiungere il massimo potenziale riproduttivo.

È importante capire che la scoperta che la scelta da parte della femmina tra i bittacidi dalle punte nere ha come risultato una variazione non casuale nella riproduzione dei maschi, indica solo il potenziale dell'evoluzione per mezzo della selezione sessuale. Per mostrare che la scelta da parte della femmina ha veramente un risultato nell'evoluzione, cioè una variazione nelle frequenze di geni all'interno di una popolazione, bisognerebbe dimostrare che i maschi scelti dalle femmine differiscono geneticamente da quelli scartati. Sembra che questo sia il caso nostro: il mio lavoro mostra che in natura i maschi che scelgono prede di grandi dimensioni lo fanno tutte le volte, e lo stesso fanno i maschi che scelgono prede di piccole dimensioni.

Uno dei modi in cui i maschi di bittacidi dalle punte nere riescono a ridurre i propri movimenti e quindi i rischi di preda-

Il numero di spermatozoi che passano dal maschio alla spermateca della femmina di bittacide dalle punte nere dipende dalla durata dell'accoppiamento. Come indica questo grafico che mostra i risultati di 66 accoppiamenti di differente durata, nessuno o pochi spermatozoi entrano nella spermateca della femmina durante i primi cinque minuti di accoppiamento. Dopo questo periodo vi è una relazione diretta tra la durata dell'accoppiamento e il numero di spermatozoi trasferiti. Per gli accoppiamenti che durano più di 20 minuti, comunque, il numero medio rimane di circa 22 000. Dal momento che solo i maschi con prede di grandi dimensioni copulano per 20 minuti o più, essi soli riescono a riprodursi.

zione da parte dei ragni consiste nel rubare una preda ad altri maschi oppure a delle coppie. Il furto della preda è un'importante componente della competizione tra i maschi per l'accoppiamento, la modalità di selezione sessuale più familiare, e la riduzione dei rischi non è l'unico vantaggio che questo comportamento conferisce. Ho osservato che un maschio che ruba la preda da offrire come dono nuziale le passa in media il 42 per cento in meno del tempo nelle sequenze preliminari dell'accoppiamento, e perciò la sua frequenza di accoppiamento è significativamente più alta di quella di un maschio che cattura da solo le prede. Viceversa, un maschio che è stato derubato della preda deve, per potersi riprodurre con successo, procurarsi un'altra offerta nuziale, e c'è una correlazione positiva tra il numero di volte che un maschio viene derubato e il tempo medio che intercorre tra due suoi accoppiamenti successivi.

Le osservazioni dei maschi marcati tra due accoppiamenti indicano pure che ogni maschio può o catturare un artropodo come preda o rubarne uno. Il modo in cui un maschio ottiene la propria preda nuziale dipende da ciò che incontra prima: un artropodo vivo oppure un maschio o una coppia in possesso di una preda. Questa strategia mista di caccia si è probabilmente evoluta perché il successo di un maschio nel furto della preda dipende dalla disponibilità di altri maschi, la quale dipende a sua volta dalle dimensioni della popolazione di bittacidi e dal rapporto numerico fra i due sessi, due fattori che variano nel tempo e nello spazio. Quando un maschio tenta di rubare una preda, può usare due diversi metodi. O vola vigorosamente sul maschio o sulla coppia che tiene una preda e si impegna in una lotta per guadagnarne il possesso oppure si comporta come una femmina per indurre un altro maschio a donargli la preda.

L'imitazione del comportamento femminile da parte del maschio è stata osservata in un'ampia varietà di animali, tra cui iene, pecore di montagna, uccelli, salamandre, pesci e primati. Il bittacide dalle punte nere mostra indiscutibilmente una vera imitazione della femmina da parte del maschio. Il successo dell'imitazione viene facilitato da diversi fattori: i maschi e le femmine hanno più o meno la stessa taglia e il medesimo colore e i genitali esterni del maschio non sono grandi ed evidenti come quelli dei maschi di molte altre specie di bittacidi. Inoltre l'imitazione delle femmine da parte dei maschi nel bittacide dalle punte nere è chiaramente adattativa: i maschi che la praticano si impossessano più rapidamente di una preda e perciò la loro frequenza di accoppiamento aumenta, per di più essi vanno meno in giro e probabilmente sono meno soggetti alla predazione da parte dei ragni che costruiscono ragnatele.

È probabile che i maschi vengano derubati della preda da altri maschi durante l'accoppiamento, ma quando un maschio sta liberando feromone è più soggetto a venir derubato da un maschio che imita una femmina, perché «si aspetta» l'arrivo di una femmina. Dal momento che è solo la femmina a rispondere al feromone, un potenziale ladro di prede viene guidato verso il maschio che libera feromoni non da stimoli olfattivi bensì da quelli visivi, ma è solo nel contesto della liberazione del feromone che l'imitazione della femmina è appropriata.

Per imitare il comportamento della femmina un maschio vola verso un altro maschio che sta liberando feromoni e abbassa le ali. Il tentativo di furto della preda è l'unica attività (eccetto il riposo notturno) in cui un maschio abbassa le ali, ma quando è una femmina ad abbassare le ali, questo costituisce un segnale rivolto al maschio, e comunica che essa è pronta a iniziare l'accoppiamento. Perciò rappresenta per il maschio un segnale per l'offerta alla femmina della preda nuziale. Quando un maschio che sta imitando una femmina abbassa le ali, comunque, il maschio che sta liberando feromoni risponde circa un terzo delle volte volando via, evidentemente perché l'imitazione non era convincente; il rimanente delle volte il maschio presenta il dono nuziale all'imitatore. Quando entrambi i maschi stanno tenendo la preda, l'imitatore inizia a nutrirsi mentre il proprietario tenta di accoppiarsi con lui. L'imitatore tiene l'addome fuori dal campo dei genitali dell'altro maschio, esattamente come fa una femmina durante il periodo iniziale di assaggio e valutazione della preda. Dopo circa due minuti il proprietario tenta di sottrarre la preda dalla presa dell'imitatore. Il 22 per cento delle volte lo sforzo non è coronato da successo e l'imitatore vola via con la preda. (Solamente una metà circa dei tentativi di rubare la preda ai maschi mentre liberano feromoni implica l'imitazione della femmina; negli esempi rimanenti il ladro effettua un tentativo diretto di afferrare la preda, tattica che ottiene successo il 14 per cento delle volte.)

In molte specie animali esiste competizione tra i maschi per quelle risorse necessarie per attirare le femmine: territorio, luoghi di nidificazione, cibo e così via, quindi le osservazioni che ho qui descritto a proposito della scelta dei maschi da parte delle femmine e della competizione dei maschi per le femmine nel bittacide dalle punte nere *Hylobittacus apicalis* possono venire applicate al comportamento di accoppiamento di una notevole varietà di animali, sia invertebrati che vertebrati. Inoltre le specie viventi dei mecotteri, il piccolo ordine di insetti a cui appartengono i bittacidi, sono i resti di un ordine che un tempo era numeroso e ampiamente diffuso. Gli studiosi dell'evoluzione degli insetti ritengono che sia stata un'antica specie di mecottero a dare origine a vari tra i più alti ordini di insetti, compresi i lepidotteri (farfalle e falene), i ditteri (mosche) e i sifonatteri (pulci). La comprensione delle pressioni evolutive che hanno plasmato il comportamento riproduttivo dei mecotteri dovrebbe migliorare la nostra comprensione del comportamento riproduttivo in generale.

Il litio e le alterazioni maniacali

Il benefico effetto dei sali di litio sui pazienti affetti da alterazioni maniacali può forse trovare una chiave interpretativa nelle modalità di attraversamento delle membrane biologiche da parte degli ioni litio

di Daniel C. Tosteson

Il termine «alterazione maniacale» si applica in senso lato a molti tipi di eccitamento dell'umore; il suo significato clinico è, invece, più specifico. L'alterazione maniacale è, infatti, uno stato patologico in cui il paziente mostra iperattività mentale e fisica, umore da euforico ad aggressivo e comportamento imprevedibile. Spesso la malattia presenta alternanza di stati di eccitamento e di stati depressivi che possono ripetersi a intervalli più o meno lunghi. I soggetti colpiti da queste forme maniaco-depressive o semplicemente da alterazione maniacale migliorano notevolmente quando sono trattati con sali di litio.

Questi sali, quando sono sciolti in acqua, si dissociano - come la maggior parte degli altri sali - in ioni, dotati rispettivamente di carica positiva e negativa. La distribuzione nel corpo umano degli ioni litio, assieme a molte loro funzioni, comporta alcune interazioni con le membrane che avvolgono le cellule e gli organelli intracellulari. In effetti, la trasmissione degli impulsi tra le varie cellule nervose cerebrali e lungo ogni singola cellula dipende dal regolare trasferimento di ioni attraverso le loro membrane ed è stato recentemente dimostrato che i pazienti affetti da forme maniacali possono presentare un difetto lungo una via che fa parte del trasporto del litio attraverso le membrane dei globuli rossi.

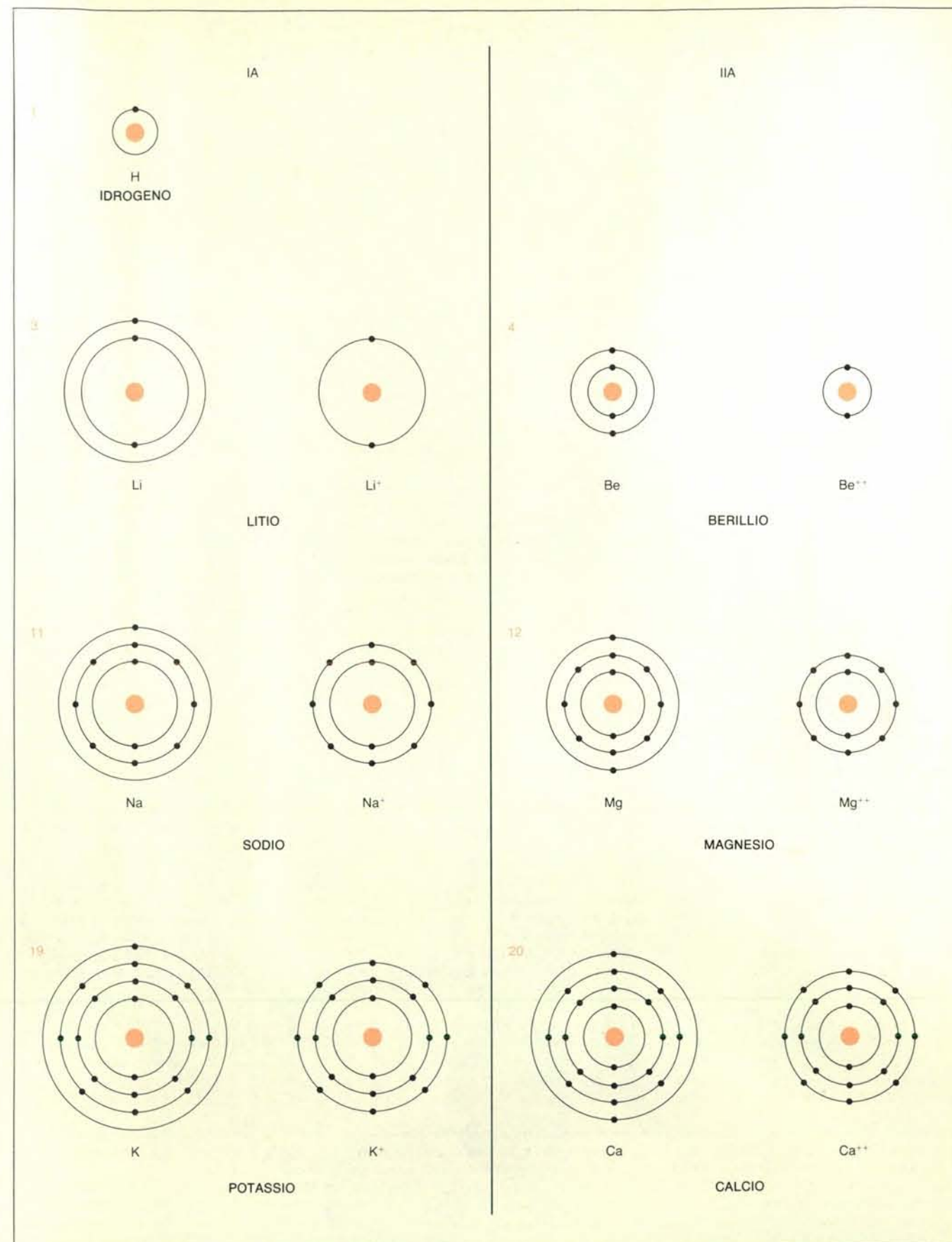
Secondo un'ipotesi piuttosto verosimile, la stessa situazione si verificherebbe nelle cellule cerebrali. Nel corso di questo articolo intendo esporre le conoscenze attuali sui rapporti tra litio, membrane e alterazioni maniacali non trascurando anche alcune ipotesi sul significato di queste nozioni per un approfondimento della comprensione della base fisico-chimica delle attività mentali umane.

La sperimentazione con il litio prese avvio nel 1949, quando venne pubblicata su «The Medical Journal of Australia» una breve relazione di J. F. J. Cade. In essa l'autore riferiva di aver notato un marcato miglioramento in pazienti affetti da alterazioni maniacali acute e croniche dopo trattamento con sali di litio. A quell'epoca, Cade lavorava da solo, come medico e psichiatra, in un piccolo ospedale per malati mentali cronici e non aveva nessuna precedente esperienza nella ricerca. Il suo interesse si concentrava sulle condizioni che davano origine a disordini affettivi del tipo degli stati maniacali e depressivi e pensava che lo stato maniacale potesse venir innescato e poi mantenuto da una sostanza rintracciabile nei liquidi organici.

Cade cominciò a cercare quell'ipotetico composto iniettando, nella cavità peritoneale di cavie, urina di pazienti affetti da alterazioni maniacali e depressive e di soggetti sani. Data la possibile presenza nell'urina, in concentrazioni elevate, dell'acido urico e dei suoi sali, Cade scelse di studiare l'effetto dell'iniezione di acido urico nei suddetti animali. E, non essendo l'acido urico e molti dei suoi sali facilmente solubili in acqua, preferì lavorare con l'urato di litio, che ha solubilità maggiore. Per controllare gli effetti del litio contrapposto all'urato, egli iniettò in altre cavie una soluzione di carbonato di litio e, con sua sorpresa, trovò che, dopo un periodo di latenza di circa due ore, gli animali cadevano in un profondo letargo e rimanevano in quella condizione un'ora o due, prima di ritornare alla loro consueta timida attività. In seguito Cade pubblicò le seguenti osservazioni: «Coloro che hanno compiuto esperimenti con le cavie, sanno in quale misura una reazione di sussulto faccia parte del temperamento di questi animali. È stato così ancora più sorprendente per lo sperimentatore trovare che, dopo iniezione di una soluzione di carbonato di litio, essi potevano essere girati sul dorso e, invece di esibire il loro abituale comportamento frenetico nel raddrizzarsi, rimanevano in quella posizione, guardandolo placidamente.»

Cade formulò l'ipotesi che l'effetto calmante del litio sulle cavie potesse essere ripetuto nell'uomo. Pertanto cominciò a studiare le conseguenze di una somministrazione orale di sali di litio in un uomo di 51 anni, che era stato ricoverato perché affetto da un sindrome maniacale cronica già da cinque anni. Dopo cinque giorni di trattamento, quel paziente cominciò a mostrare un deciso miglioramento, ed entro due mesi fu in grado di lasciare l'ospedale e di riprendere il lavoro, pur continuando il trattamento con il litio.

Dopo il lavoro iniziale di Cade, il ruolo della terapia a base di litio nelle alterazioni mentali fu analizzato con il massimo impegno da Mogens Schou e collaboratori presso l'Università di Aarhus, in Danimarca. A partire dall'inizio degli anni sessanta, il litio venne accettato come farmaco di elezione nella cura delle affezioni maniacali sia in Australia sia nella massima parte dei paesi europei. Dovettero passare, invece, altri dieci anni prima che esso fosse accettato anche negli Stati Uniti. Una ragione di questo grande ritardo fu la estrema cautela mostrata da molti medici americani nei riguardi del litio, a causa della disastrosa esperienza fatta con esso negli anni quaranta quando veniva sostituito al sodio nella cura delle insufficienze cardiache congestizie. Parecchi pazienti erano morti prima che la tossicità del litio, che è particolarmente pronunciata in coloro che hanno una carenza di sodio, divenisse manifesta. Contribuivano al ritardo anche fattori economici. Per la semplicità chimica del litio e la sua pronta disponibilità, le società farmaceutiche non potevano ricavare notevoli guadagni producendo questo farmaco e quindi esse avevano uno scarso incentivo a produrre sali di litio e a eseguire tutte le prove richieste dalla Food and Drug Administration statunitènse prima di introdurli nella pratica clinica.

Nella scala periodica degli elementi, il litio è per leggerezza il terzo in ordine decrescente, dopo l'idrogeno e l'elio. Il suo atomo ha tre protoni (cariche positive) nel nucleo e tre elettroni (cariche negative) che orbitano attorno al nucleo. Uno di questi elettroni è solo nell'orbitale più esterno e pertanto viene ceduto facilmente interagendo con altri atomi. Senza questo elettrone, l'atomo di litio ha tre cariche positive nel nucleo e due negative nel «guscio» costituito dagli orbitali con

Lo ione litio viene qui rappresentato insieme all'idrogeno e agli altri ioni in grado di attraversare le membrane cellulari. Gli elementi sono elencati secondo la loro posizione nei gruppi IA e IIA della tavola periodica degli elementi, oltre che in base alle loro dimensioni relative. I numerali in colore indicano il numero di protoni nel nucleo di ogni atomo; i punti neri rappresentano gli elettroni. Un atomo di litio ha due elettroni nel primo orbitale e uno soltanto nel secondo (il più esterno): un totale di tre cariche negative, che si bilanciano con le tre cariche positive del nucleo. L'elettrone esterno viene facilmente ceduto, per cui l'atomo diventa un catione, cioè uno ione dotato di carica positiva (Li^+). Parecchi pazienti affetti da forme maniacali possono essere curati con sali di litio i quali, nei liquidi organici, si dissociano in ioni.

gli elettroni. Pertanto è un catione, cioè uno ione dotato di carica positiva, che interagisce con gli anioni o ioni dotati di carica negativa. Nella sintesi degli elementi più pesanti a partire dai più leggeri, che si svolge nell'universo, il litio non si trova sulla via principale: pertanto, sulla Terra, in confronto ad altri elementi chimici che gli sono affini - come il sodio, il potassio, il magnesio e il calcio - è presente solo in piccole quantità. Malgrado questo, ha un'ampia distribuzione nei minerali, soprattutto nei silicati.

Per quanto riguarda la chimica fisica, sorprende la grande similitudine tra sodio, litio e potassio e, proprio per questa proprietà, i tre elementi sono stati classificati tutti nel gruppo IA della tavola periodica. Per la fisiologia, invece, assumono particolare significato proprio le differenze che li contraddistinguono.

La capacità delle membrane della cellula vivente di generare e propagare segnali elettrici dipende dalla capacità di queste membrane di discriminare tra il sodio e il potassio. I due elementi devono anche essere riconosciuti dalla macchina molecolare che pompa il sodio fuori e il potassio dentro le cellule nervose e muscolari, caricando così la «batteria» che emette gli impulsi elettrici. Gli effetti del litio negli organismi viventi, compresi quei soggetti umani con turbe affettive come le alterazioni maniacali, dipendono dalle differenze esistenti tra gli ioni di questo elemento e gli ioni di altri membri dei gruppi IA e IIA, che sono presenti nelle cellule, in particolar modo il sodio, il potassio, il calcio e il magnesio.

La distribuzione degli ioni negli organismi viventi e le velocità con cui essi si spostano sono in gran parte regolate dalle membrane che separano l'interno e l'esterno delle cellule e degli organelli intracellulari come i mitocondri. Come è noto, le membrane di questo tipo sono costituite da un doppio strato di molecole di fosfolipidi e di colesterolo, in cui sono incluse molecole proteiche specializzate. Il loro spessore, comprese le proteine, è di circa 10 nanometri (cioè dieci milionesimi di millimetro).

La parete di queste membrane biologiche costituita dal doppio strato lipidico permette senza difficoltà il passaggio di piccole molecole, prive di carica e relativamente solubili nei lipidi, come sono quelle dell'anidride carbonica e dell'ossigeno, ma non il passaggio degli ioni dotati di carica elettrica e solubili in acqua, come quelli del sodio, del potassio e del litio. La maggior parte del trasporto degli ioni, se non tutto, attraverso una membrana biologica è quindi mediato da proteine specifiche incluse nel doppio strato lipidico. Quattro tipi di trasporto meritano di essere menzionati. Innanzitutto, vi possono essere canali proteici che consentono il movimento selettivo di vari ioni lungo i loro rispettivi gradienti di concentrazione e di potenziale elettrico. Un esempio di questo tipo è fornito dai canali per il sodio e per il potassio, che dipendono dal tempo e dalla differenza di potenziale e sono responsabili della propagazione del potenziale d'azione (l'impulso nervoso) nelle fibre nervose e muscolari. Un altro esempio è dato dai canali, attivati dall'acetilcolina, che innescano il potenziale d'azione nelle fibre muscolari.

Un secondo tipo di trasporto è consentito da «pompe» che trasformano l'energia dei legami chimici - qual è quella contenuta nel legame altamente energetico del pirofosfato nell'ATP - nel lavoro necessario per spostare gli ioni contro i gradienti di concentrazione e di potenziale elettrico. Si citano, tra i vari esempi, la pompa del sodio-potassio, presente nella membrana esterna della maggior parte delle cellule e le pompe del calcio presenti nelle membrane dei mitocondri e del reticolo sarcoplasmatico.

Un terzo tipo di trasporto ionico è stato scoperto da Hans H. Ussing dell'Università di Copenaghen, che l'ha chiamato diffusione per scambio. Questo processo permette uno scambio, ma nessun movimento netto di ioni attraverso la membrana. Esso venne osservato per la prima volta da Ussing e collaboratori nel muscolo di rana come scambio tra sodio e sodio. L'individuare uno scambio che comporti un'unica specie di ioni (nell'osservazione di Ussing il sodio) è possibile solo grazie a un tracciante, come un isotopo radioattivo dello ione trasportato. Se, invece, in un sistema di questo tipo partecipano due differenti tipi di ioni (come il sodio e il litio), lo scambio procede fino a quando il rapporto tra le concentrazioni delle due specie ioniche è identico in ambedue le soluzioni (interna ed esterna), che bagnano la membrana. Gli ioni cloro e bicarbonato si scambiano attraverso le membrane dei globuli rossi umani proprio grazie a questo tipo di meccanismo.

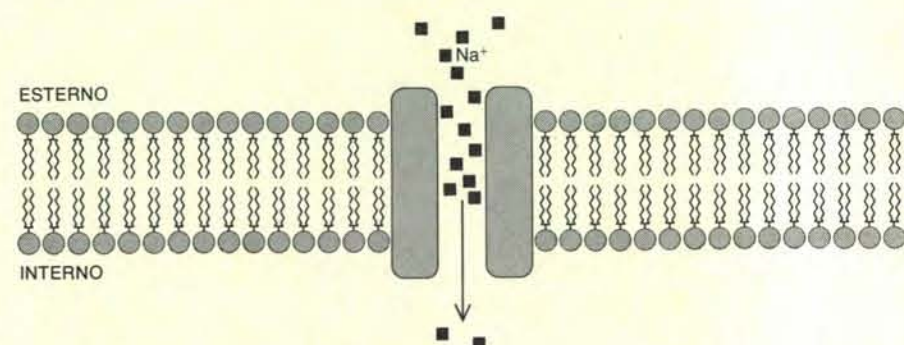
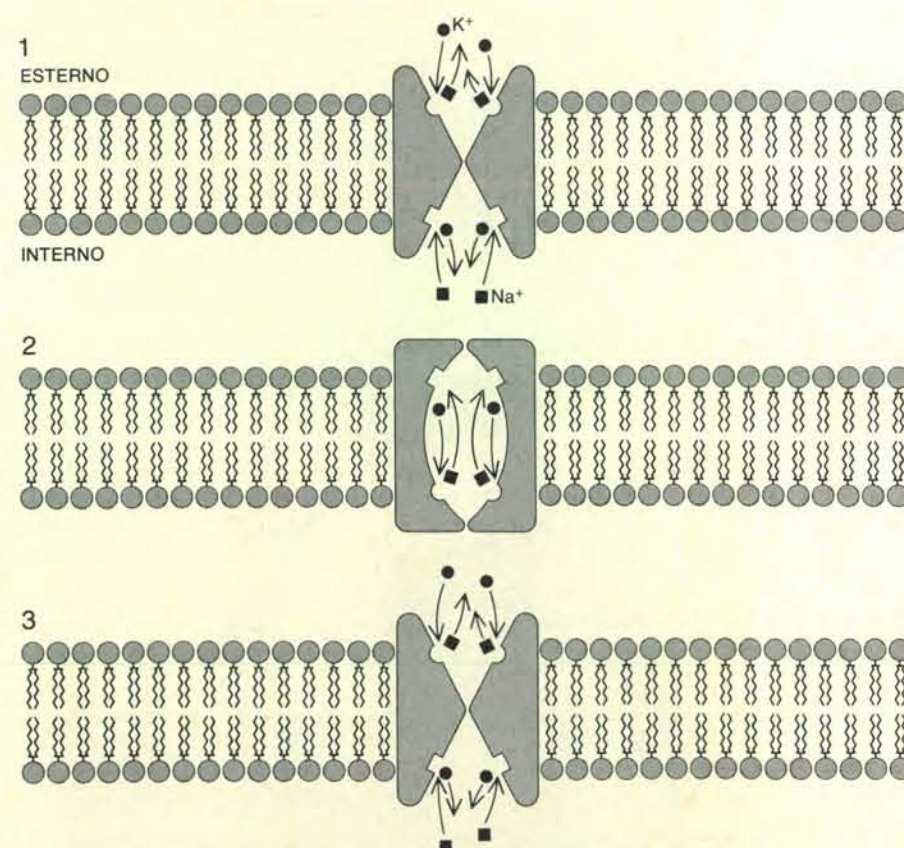
Viene qui illustrata schematicamente, secondo le attuali concezioni, la struttura della membrana cellulare. Essa è costituita da un doppio strato di molecole lipidiche (grassi) con le teste idrofile rivolte verso l'esterno e le code idrofobe orientate verso l'interno. Associate alla membrana vi sono molecole proteiche - qui rappresentate come corpi più grossi - che possono attraversarla completamente o esservi solo parzialmente immerse. Gli spostamenti degli ioni verso l'interno o l'esterno della membrana avvengono perlopiù lungo canali a livello delle varie proteine.

Un quarto tipo di movimento molecolare attraverso le membrane biologiche è il cotrasporto, che consente lo spostamento simultaneo di due sostanze. Ne sono esempi il movimento del sodio e degli zuccheri o del sodio e degli amminoacidi nelle cellule dell'epitelio intestinale come parte del processo di assorbimento e il trasporto accoppiato del sodio e del potassio nei globuli rossi degli uccelli e in alcune altre cellule. Il litio può sostituire il sodio in tutte queste vie. Nelle condizioni che si osservano nel sangue di pazienti trattati con litio, invece, solo una via - lo scambio sodio-litio o controtrasporto - fa spostare il litio attraverso la membrana dei globuli rossi verso l'esterno e contro il suo gradiente di concentrazione.

Oltre a regolare la distribuzione degli ioni e di molecole di piccole dimensioni, le membrane svolgono molte altre importanti funzioni cellulari, tra le quali l'assorbimento e l'eliminazione di grosse molecole e di piccoli volumi di liquido, il riconoscimento di molecole extracellulari - come gli ormoni, gli antigeni e gli anticorpi - e di altre superfici cellulari nello sviluppo embrionale e nel processo di riparazione di organi; infine, l'ordinamento della sequenza di varie reazioni biochimiche.

Alcune connessioni tra il litio, le membrane e le affezioni maniacali cominciarono a chiarirsi quando Joe Mendels e Alan Frazer della University of Pennsylvania School of Medicine osservarono che il rapporto o quoziente di litio (cioè il valore della concentrazione del litio all'interno dei globuli rossi diviso per il valore della concentrazione del litio nel plasma sanguigno all'esterno dei globuli) variava considerevolmente tra i pazienti che erano stati sottoposti a una terapia a base di litio. Mendels e Frazer trovarono anche che il valore medio in un gruppo di pazienti con turbe affettive e sensibili al litio era decisamente più alto di quello riscontrato in altri gruppi di pazienti e in soggetti normali, usati come controlli.

Mendels e Frazer hanno anche riferito che il quoziente di litio era più elevato nei globuli rossi di pecora con un'alta concentrazione di sodio che non in quelli con bassa concentrazione. Questo risultato mi ha fatto intuire l'esistenza di un sistema di controtrasporto, riguardante gli ioni sodio e litio, nelle membrane dei globuli rossi umani. Alcuni esperimenti che Mark Haas, James M. Schooler e io eseguimmo nel 1975 alla Duke University School of Medicine hanno confermato l'ipotesi. Un po' di tempo dopo, in collaborazione con G. N. Pandey e John Davis della Duke University, con Balasz Sakardi e R. B. Gunn dell'Università di Chicago e con J. Funder e J. O. Wieth dell'Università di Copenaghen, abbiamo analizzato i processi di trasporto che determinano la distribuzione del litio tra globuli rossi umani e plasma sanguigno. Simultaneamente, ma in maniera del tutto indipendente, Jochen Duhm e collaboratori della Facoltà di medicina dell'Università di Monaco di Baviera hanno effettuato alcune analoghe osservazioni, poi confermate ed estese da Barbara E. Ehrlich e Jared M. Diamond dell'Università della California.

La concentrazione del litio nel plasma sanguigno di pazienti in cura per turbe affettive è di circa una millimole per litro (cioè è una concentrazione millimolare). Essa è il risultato di un equilibrio dinamico tra il passaggio verso l'interno degli ioni litio lungo il gradiente di concentrazione e il controtrasporto verso l'esterno contro tale gradiente. Il movimento con-

Il canale proteico, attraverso il quale ioni come il sodio, il potassio e il litio si possono spostare lungo il gradiente di concentrazione, interessa una proteina che penetra nello spessore della membrana. Il movimento lungo il gradiente, che può essere orientato o verso l'interno della cellula o verso l'esterno, secondo le differenze nella concentrazione ionica e nel potenziale elettrico che si stabiliscono sui due lati della membrana, non richiede energia metabolica. In questo schema, in particolare, gli ioni sodio si stanno spostando verso l'interno della membrana.

La pompa ionica fornisce un altro mezzo per il trasporto di ioni dall'interno della cellula verso l'esterno e viceversa. Essa trasforma l'energia dei legami chimici nel lavoro necessario a far spostare gli ioni contro i loro gradienti di potenziale elettrochimico. Un ipotetico meccanismo per la pompa sodio-potassio, che consta di due stati e di quattro siti, comincia *(1)* con una proteina che penetra nella membrana e che si allarga a imbuto all'interno e all'esterno della membrana stessa. Sul lato esterno, uno ione potassio proveniente dal plasma sanguigno si sposta in direzione di un sito di legame, lasciato libero da uno ione sodio; uno scambio inverso ha luogo all'interno. La configurazione del canale proteico si modifica e passa a uno stato chiuso *(2)*, la selettività dei due siti cambia ugualmente e il potassio, che si trova sul sito rivolto verso l'esterno, si scambia con il sodio, situato nel sito rivolto verso l'interno. Si reintegrano la prima configurazione e la selettività dei siti *(3)* e il ciclo si chiude. I cambiamenti di forma e di selettività sono alimentati dall'energia che si libera grazie al processo di idrolisi dell'adenosintrifosfato (ATP). Non esiste mai una via aperta che permetta agli ioni di spostarsi lungo l'intera distanza attraverso la membrana.

tro gradiente è alimentato non dalla liberazione dell'energia immagazzinata nei legami chimici (come nell'ATP), ma da una dissipazione del gradiente di concentrazione del sodio, diretto verso l'interno. Il sodio che entra nella cellula attraverso il sistema di scambio litio-sodio viene riportato all'esterno dalla pompa sodio-potassio, che è alimentata dall'ATP. Il sistema del controtrasporto ha, quindi, l'effetto di accoppiare il litio alla pompa sodio-potassio. Il quoziente di litio all'equilibrio si avvicina a quello di 0,08 del sodio, quando il sistema di controtrasporto litio-sodio è molto attivo e vi è solo poca o nessuna dispersione. Quando la via lungo la quale avviene la dispersione è grande, mentre quella per il controtrasporto è piccola, il quoziente di litio si avvicina a 1,2, che è il valore per i cationi passivi con un'unica carica elettrica, quando il potenziale elettrico all'interno dei globuli rossi è di -10 millivolt rispetto al plasma esterno. È questo il caso dei globuli rossi umani. Nella maggior parte dei soggetti il quoziente di litio è compreso tra 0,2 e 0,6.

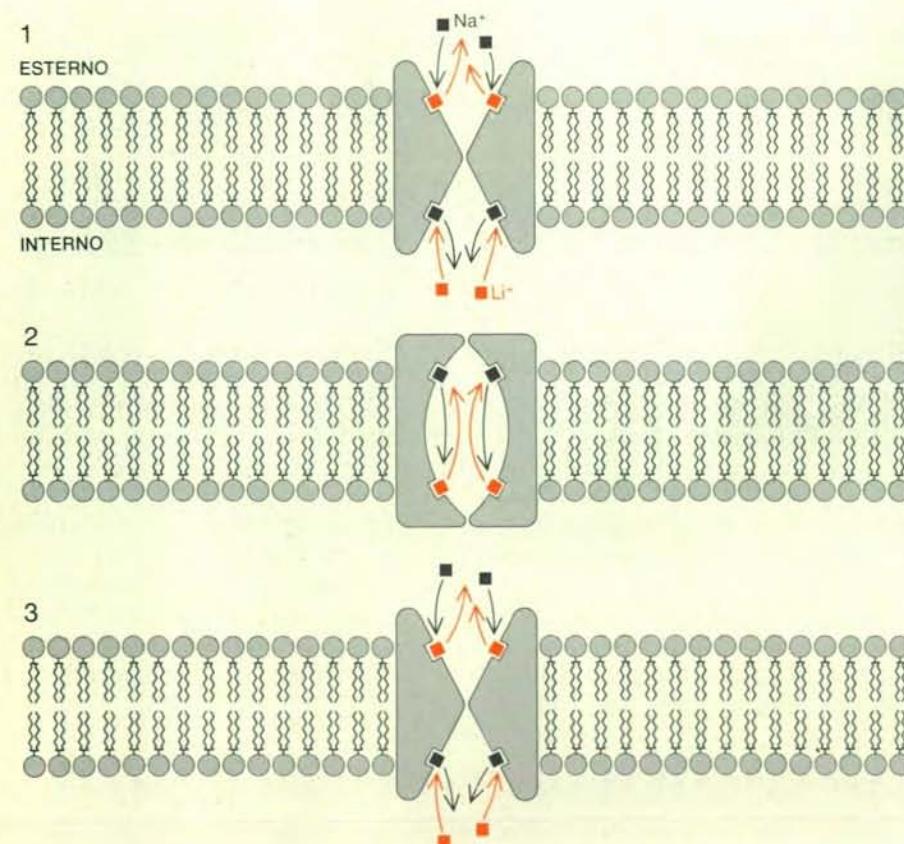
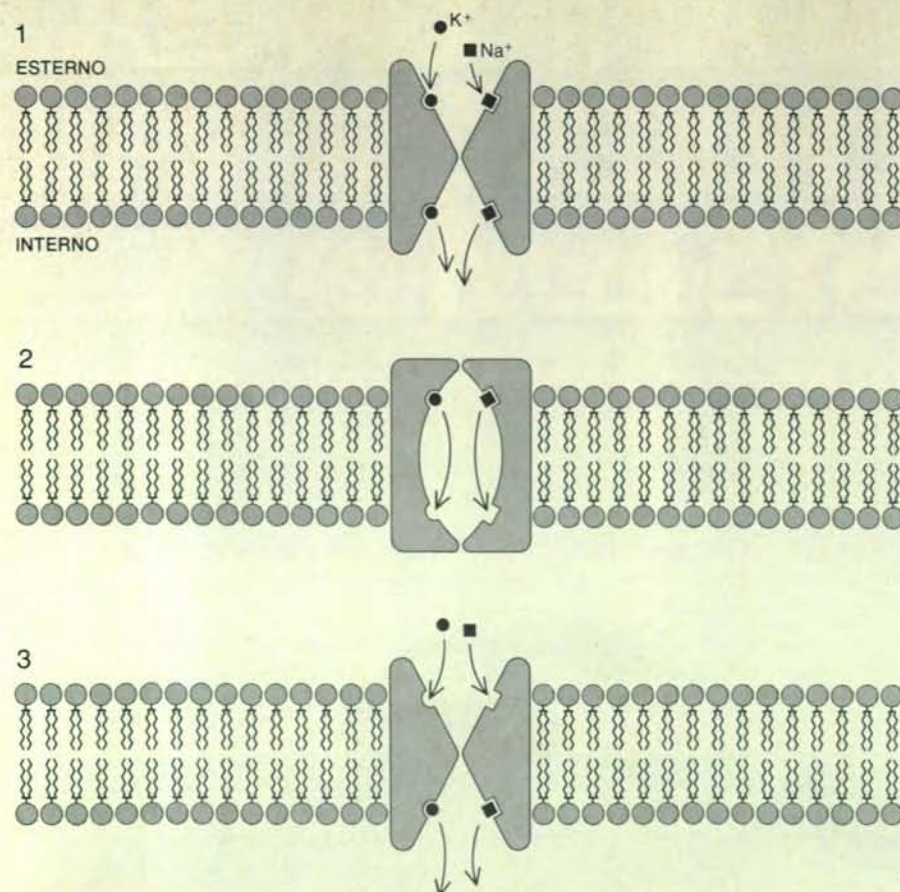
Il sistema di controtrasporto realizza uno scambio 1:1 tra sodio e sodio, o tra litio e sodio. A seconda del gradiente di sodio, esso sposta il litio contro gradiente o verso l'interno o verso l'esterno dei globuli rossi. Dato che, nel corpo, la concentrazione del sodio all'interno delle cellule è quasi sempre molto più bassa della concentrazione all'esterno, il sistema di controtrasporto dà origine a un movimento netto del litio verso l'esterno ogniqualvolta il quoziente di litio è più alto di quello del sodio. L'affinità del sistema per il litio è di circa 20 volte superiore a quella per il sodio, come dimostra il fatto che la concentrazione intracellulare necessaria per una stimolazione del controtrasporto a metà del valore massimo è 0,5 millimolare per il litio e 10 millimolare per il sodio.

Il sistema di controtrasporto non accetta il potassio, l'ammonio, la colina, il calcio, il magnesio o altri cationi finora provati. Non viene inibito dalla uabaina, che blocca in maniera specifica la pompa sodio-potassio alimentata dall'ATP. Viene inibito, invece, dalla floretina, un ben noto inibitore dei sistemi di trasporto per gli zuccheri e gli anioni. Non richiede l'ATP o la glicolisi (cioè il metabolismo del glucosio che fornisce energia ai globuli rossi del sangue).

La velocità massima di trasporto del sistema di controtrasporto litio-sodio varia in maniera considerevole da individuo a individuo. L'ambito di variazione è tra zero e 0,7 millimoli per litro di globuli rossi all'ora. Pandey e Duhm con i rispettivi collaboratori hanno dimostrato che simili valori sono correlati con la variazione individuale nei confronti del quoziente di litio nei globuli rossi, quoziente che viene misurato sia nei pazienti trattati con litio sia in vitro (con globuli rossi incubati per 24 ore in un mezzo simile al plasma, contenente 1,5 millimoli per litro di litio).

Misure eseguite su gemelli e individui imparentati tra loro presentanti turbe affettive e su soggetti normali di controllo fanno pensare che le variazioni individuali nella velocità massima del controtrasporto litio-sodio siano determinate (almeno in parte) geneticamente.

La penetrazione del litio nei globuli rossi umani lungo il gradiente di concentrazione può avvenire seguendo due vie. Una non è inibita da alcuno dei composti finora provati e non sembra interessare altri ioni o molecole prive di carica elettrica. L'altra è una via interessante, scoperta per la prima volta da Funder e Wieth. Gli ioni litio (e in minor misura quelli sodio) possono reagire con il carbonato, un anione dotato di una duplice carica elettrica,formando gli ioni carbonato di litio e carbonato di sodio, dotati di un'unica carica elettrica. Questi anioni possono essere trasportati lungo un sistema di scambio di anioni con carica singola, molto rapido, il quale normalmente fa passare attraverso la membrana del globulo rosso gli ioni cloro e bicarbonato. Anche se le concentrazioni del carbonato di litio e del carbonato di sodio, e pertanto le loro velocità di trasporto, sono sempre basse in confronto con altri ioni che si spostano nel sistema di scambio anionico, la velocità di spostamento del litio lungo questa via è notevole rispetto ad altre vie. In pazienti con concentrazioni normali di ioni cloro e bicarbonato nel plasma, circa la metà del litio che penetra all'interno dei globuli rossi passa attraverso la via del carbonato.

Lo spostamento del litio attraverso la membrana dei globuli rossi di una persona che viene curata con questo elemento avviene in primo luogo lungo la via sopradescritta e il sistema di controtrasporto litio-sodio. In certe condizioni sperimentali, invece, si hanno altre vie che consentono allo ione litio di attraversare la membrana del globulo rosso. Per esempio, esso entra in competizione sia con i siti del sodio che con quelli del potassio nella pompa sodio-potassio. Quando i due elementi non sono presenti all'esterno della cellula, il litio sostituisce il potassio e viene attivamente trasportato all'interno del globulo rosso; quando i due elementi, invece, mancano all'interno della cellula, il litio sostituisce il sodio e

Rappresentazione schematica di scambio ionico o di controtrasporto di ioni che si ritiene contribuisca allo spostamento degli ioni litio verso l'esterno dei globuli rossi nella terapia delle alterazioni maniacali. Esso effettua normalmente uno scambio 1:1 del sodio con il sodio, ma può anche scambiare il sodio con il litio. Il modello ipotetico a due stati e quattro siti, qui illustrato, assomiglia al meccanismo della pompa illustrata nella figura in basso della pagina precedente. Nello stato di apertura *(1)*, i due siti possono scambiarsi gli ioni sodio e litio presenti nelle soluzioni che lambiscono le superfici interna ed esterna della membrana. Nello stato di chiusura *(2)* i due siti possono ancora scambiarsi gli ioni tra loro, ma non con le due soluzioni. Una reintegrazione dello stato iniziale completa il ciclo *(3)*. La selettività dei siti è la stessa nei due stati sopra menzionati. Nel corpo la concentrazione del sodio è minore all'interno delle cellule rispetto all'esterno; il sistema di controtrasporto produce un movimento netto di litio verso l'esterno quando il rapporto tra la concentrazione nella cellula e nel plasma è maggiore per il litio che per il sodio. Il sistema di controtrasporto, difettoso in certi pazienti con alterazioni maniacali, dà origine a un rapporto, tra le concentrazioni di litio nella cellula e nel plasma, che è superiore a quello normale.

Il sistema di cotrasporto fa muovere simultaneamente due tipi di ioni verso l'interno o verso l'esterno della membrana. In questo modello ipotetico a due stati e quattro siti, gli ioni sodio e potassio si spostano lungo il gradiente di concentrazione verso l'interno di un globulo rosso. Con la proteina che penetra nella membrana aperta verso le superfici esterna e interna della membrana stessa *(1)*, uno ione sodio e uno ione potassio si spostano dal plasma sanguigno ai siti di legame vicino alla apertura della proteina rivolta verso l'esterno, mentre due ioni analoghi si spostano dai siti di legame interni all'interno della cellula. La configurazione si modifica in uno stato di chiusura *(2)* e i due ioni dal plasma sanguigno si spostano verso i siti di legame sulla faccia interna. Poi entrano nella cellula *(3)* e il ciclo ricomincia. Contrariamente ai modelli per la pompa sodio-magnesio e per lo scambio sodio-sodio, nello stato di chiusura di questo sistema i siti possono essere vuoti.

viene attivamente trasportato all'esterno del globulo rosso. Dato che la pompa ha una maggiore affinità per il sodio e il potassio che per il litio, non è molta la quantità di litio che viene trasportata nel corpo per questa via, perché il sodio e il potassio sono sempre presenti su ambedue i lati della membrana (e in concentrazioni ben più elevate di quella del litio). Abbiamo dimostrato di recente che il litio si può sostituire al sodio nel cotrasporto del sodio e del potassio nei globuli rossi umani. Altre vie possono essere importanti per il trasporto in altri tipi di cellule. Halvor N. Christensen e collaboratori della University of Michigan Medical School hanno dimostrato recentemente che, in alcune cellule tumorali di topo, il litio si sposta insieme a determinati amminoacidi.

Quali sono, allora, le connessioni tra il litio, le membrane e la sindrome maniacale? Innanzitutto, il controtrasporto litio-sodio nei globuli rossi sembra avvenire a velocità inferiore negli individui affetti da alterazioni maniacali o disturbi maniaco-depressivi rispetto a quanto avviene negli individui normali o con altri tipi di turbe mentali. Studi familiari effettuati su persone con e senza turbe affettive da parte di Elizabeth Dorus e collaboratori della Medical School della Università di Chicago fanno pensare che un sistema di controtrasporto litio-sodio carente compaia più spesso nei parenti di primo grado (genitori o figli) di pazienti affetti da alterazioni maniacali o disordini maniaco-depressivi che nelle altre persone. Sulla base di questo risultato, si può sostenere l'ipotesi che un controtrasporto litio-sodio insufficiente sia un'alterazione ereditaria e che i geni responsabili per la sua espressione siano anche interessati in una predisposizione alle sindromi maniacali e maniaco-depressive. Poiché la maggior parte di coloro che soffrono di turbe affettive non hanno un controtrasporto litio-sodio deficitario, è chiaro che una simile ipotesi si può applicare solo a un sottogruppo di pazienti affetti da sindromi maniacali e maniaco-depressive. Ciò non sorprende, perché è ormai dimostrato che una predisposizione ereditaria a turbe affettive richiede l'interazione di molti geni. Studi più ampi sono necessari per chiarire il rapporto tra ereditarietà del sistema di controtrasporto litio-sodio e una predisposizione alle alterazioni maniacali e maniaco-depressive.

Se è giusta l'ipotesi secondo cui esiste una correlazione tra controtrasporto litio-sodio deficitario nei globuli rossi e incidenza delle alterazioni maniacali, quale potrebbe essere la connessione tra questi fenomeni così diversi? È improbabile che ciò che avviene a livello delle membrane dei globuli rossi possa influire sull'umore di una persona. È possibile, invece, che i geni che codificano per il sistema di controtrasporto litio-sodio esprimano anche un analogo sistema in altre cellule, tra cui le cellule cerebrali.

La selettività ionica della via di scambio per il sodio in queste cellule cerebrali potrebbe essere leggermente diversa, permettendo lo scambio del sodio con il calcio o con vari neurotrasmettitori costituiti da cationi con una carica elettrica singola. È stato dimostrato che il litio interagisce con il sistema di scambio sodio-calcio nelle membrane plasmatiche delle cellule nervose e muscolari e con i sistemi che, nel cervello, fissano i neurotrasmettitori. Un difetto ereditario in tale sistema di trasporto nel cervello potrebbe contribuire a determinare quell'anomalia mentale nota come sindrome maniacale.

Devo sottolineare che qualunque relazione venga trovata tra la determinazione genetica del controtrasporto litio-sodio e le alterazioni maniacali (o il meccanismo dell'attività terapeutica del litio) è, allo stato attuale delle conoscenze, puramente speculativa. Non sappiamo ancora, infatti, se le alterazioni maniacali sono la manifestazione di un anormale funzionamento di un piccolo gruppo di cellule cerebrali oppure se sono il segno di una disfunzione più generalizzata. Non è neppure conosciuto il meccanismo attraverso cui il litio agisce sulle alterazioni maniacali: modifica la funzione di un piccolo gruppo di cellule cerebrali o di molte di esse? L'azione terapeutica del litio non sembra correlata alle variazioni individuali nella velocità di controtrasporto litio-sodio nei globuli rossi; è altrettanto efficace in pazienti dotati di un sistema di controtrasporto attivo e in pazienti in cui tale controtrasporto è carente.

Date le similitudini fisico-chimiche tra gli ioni litio e i cationi sodio e potassio, normalmente presenti e importanti dal punto di vista fisiologico, non sorprende il fatto che il litio a concentrazioni sufficientemente alte interagisca praticamente con tutti i componenti dell'organismo: proteine (compresi gli enzimi), acidi nucleici e gruppi atomici dotati di carica presenti sui lipidi e sui carboidrati. Non sorprende neppure la quantità di effetti biologici riportata in casi di somministrazione di litio.

Anche se la nostra attenzione è rivolta esclusivamente agli effetti in campo terapeutico della concentrazione di ioni litio, si possono trarre interessanti conclusioni sul meccanismo di azione dello ione in

generale sulle alterazioni maniacali. Tra le più promettenti, dal mio punto di vista, ricorderò: *1*) i suoi effetti sulla sintesi delle prostaglandine: *2*) i suoi effetti sulla produzione di globuli bianchi; *3*) i suoi effetti sul trasporto dei neurotrasmettitori e sulla sensibilità dei recettori a essi; *4*) la sua inibizione del sistema di trasporto della colina durante la maturazione dei globuli rossi umani; *5*) la sua efficacia nel far aumentare la concentrazione della colina nei globuli rossi, come hanno riferito di recente Ehrlich e Diamond; *6*) la sua inibizione dell'attivazione dell'adenilciclasi (l'enzima che catalizza l'azione del messaggero intracellulare adenosinmonofosfato, o AMP) a opera delle catecolammine, dell'ormone tireotropo e dell'ormone antidiuretico.

Tenendo conto dei numerosi siti d'azione del litio, non sorprende che esso possa avere molti effetti tossici. L'ambito di sicurezza del dosaggio terapeutico è ristretto. Tra le complicazioni più gravi che intervengono nel corso di una terapia con il litio si annovera quella che può impedire ai reni di reagire all'ormone antidiuretico, provocando così un'alterazione nota come diabete insipido.

Viene qui illustrata schematicamente la presenza del litio nel sangue umano, quando un paziente viene trattato con uno dei suoi sali. Gli ioni che figurano nel processo sono il litio *(quadrati in colore)*, il sodio *(quadrati neri)*, il potassio *(dischi neri)*, il carbonato di litio *(triangoli colorati)*, il bicarbonato *(triangoli neri)* e il cloro *(triangoli vuoti)*. I simboli chimici di questi cationi e anioni sono rispettivamente Li^+, Na^+ e K^+, $LiCO_3^-$, HCO_3^- e Cl^-. In un paziente curato con litio, la concentrazione di Li^+ nel plasma sanguigno *(in colore chiaro)* viene mantenuta a circa 1 millimole per litro. La concentrazione all'interno delle cellule è mantenuta, invece, a un livello più basso (circa 0,3 millimoli per litro) dall'attività di un sistema di controtrasporto litio-sodio. Quando si raggiunge uno stato di equilibrio, il movimento del litio verso l'esterno attraverso il sistema di controtrasporto è bilanciato da un movimento verso l'interno degli ioni attraverso due vie: un'infiltrazione e uno scambio anionico, che cede il $LiCO_3^-$ al posto di HCO_3^- o di Cl^-. Parecchi movimenti attraverso la membrana sono illustrati nella figura in alto; in quella in basso, invece, viene illustrata una condizione di equilibrio in un momento qualsiasi. In ogni rappresentazione, sono indicate le proporzioni relative dei vari cationi e anioni nella cellula e nel plasma.

Che significato può avere per il futuro della medicina questo capitolo incompleto sul litio, le membrane e le alterazioni maniacali? Esso è un ulteriore esempio della crescente precisione con cui oggi è possibile valutare le dimensioni fisico-chimiche di una malattia. Sono trascorsi solo trent'anni dal primo risultato ottenuto in questo campo; la descrizione della anemia falciforme come malattia molecolare da parte di Linus Pauling e Harvey A. Itano del California Institute of Technology. Da allora, ricercatori di numerosi laboratori hanno effettuato esperimenti che hanno permesso di identificare non solo le esatte sostituzioni di amminoacidi che si verificano nell'anemia falciforme e altre forme di emoglobina anomala, ma anche l'identità chimica dei geni che dirigono la sintesi di questa essenziale proteina.

Analoghe ipotesi sono state formulate per spiegare il ruolo degli eventi molecolari nello sviluppo di alterazioni complesse come l'aterosclerosi e il diabete. Per esempio, Joseph L. Goldstein e Michael S. Brown del Health Science Center dell'Università del Texas a Dallas hanno identificato i recettori responsabili dell'assunzione da parte di fibroblasti umani di lipoproteine a bassa densità, che incorporano il colesterolo, e hanno dimostrato che tali recettori presentano, in alcuni pazienti con un alto livello di colesterolo nel sangue, un difetto ereditario. Si può quindi avere una predisposizione all'aterosclerosi.

Analogamente ci si rende conto oggi che il diabete mellito (la comune alterazione metabolica dovuta a concentrazioni troppo elevate, rispetto alla norma, di zucchero nelle urine e nel plasma sanguigno) non ha una, ma molte cause, che possono comportare difetti nella sintesi e nella messa a punto dell'insulina o nel legame che essa stabilisce con i recettori presenti sulle cellule e che spiega la sua azione sulle cellule bersaglio. Difetti del genere possono essere ereditari oppure causati da un attacco virale al sistema.

Tutto ciò fa pensare che l'attuale classificazione delle malattie in categorie come l'ipertensione, il diabete, il cancro, le alterazioni maniacali sia troppo grossolana. Le classificazioni così generiche stanno, in effetti, per essere soppiantate dall'individuazione molto più precisa dei difetti molecolari, che sono alla base delle manifestazioni generali di un'alterazione fisiologica. Ci si rende anche conto del collegamento tra tali difetti molecolari, l'interazione di geni e condizioni specifiche associate con l'ambiente. Strategie razionali ed efficaci per la prevenzione e il controllo di malattie complesse saranno messe a punto sulla base di questo tipo di interpretazione.

GIOCHI MATEMATICI

di Martin Gardner

Lavinia deve trovare una stanza sul Viale dell'Università e altri problemi geometrici

La maggior parte dei brevi problemi che compongono la rubrica di questo mese hanno a che fare con la geometria, intendendo la parola in senso ampio così da farvi rientrare la geometria combinatoria, la topologia e la teoria dei grafi. La soluzione di tutti i problemi verrà data nella mia prossima rubrica, mentre i commenti alle lettere che mi giungeranno dai lettori appariranno come minimo tra quattro mesi.

1. Lavinia cerca una stanza. La linea della figura in basso rappresenta il Viale dell'Università di una piccola città universitaria dove Lavinia studia. I punti contrassegnati con le lettere da *A* a *K* sono edifici prospicienti il viale in cui vivono gli 11 migliori amici di Lavinia.

Lavinia ha vissuto finora con i suoi genitori in una città vicina, ma ora desidera trasferirsi in Viale dell'Università. Le piacerebbe una stanza o un appartamento in posizione *L* tale per cui la somma delle distanze dai suoi 11 amici fosse la minima possibile. Ammesso che vi sia una casa disponibile nella giusta posizione, specificate dove Lavinia dovrebbe vivere e dimostrate che la somma delle distanze dalle altre posizioni è la minima possibile.

2. Solidi specularmente simmetrici. Nelle figure piane, un asse di simmetria è una linea retta che divide la figura in due metà congruenti che sono immagine speculare una dell'altra. Nelle carte da gioco, per esempio, i cuori hanno un asse di simmetria e così pure le picche e i fiori, mentre i quadri ne hanno due. Un quadrato ha quattro assi di simmetria, una stella regolare a cinque punte ne ha cinque e un cerchio ne ha un numero infinito. Una svastica o un simbolo yin-yiang non hanno nessun asse di simmetria.

Se una figura piana ha almeno un asse di simmetria, si dice che è sovrapponibile alla sua immagine speculare nel senso seguente. Se immaginate di vedere la figura in uno specchio verticale con lo specchio che poggi su un piano orizzontale, potete pensare di far scivolare la figura nello specchio e, se necessario, di ruotarla sul piano in modo da farla coincidere con la sua immagine speculare. Non è ammesso rovesciare la figura sul piano perché questo richiederebbe il farla ruotare in una terza dimensione.

Un piano di simmetria è un piano che taglia una figura solida in metà congruenti, ciascuna delle quali è una riflessione speculare dell'altra. Una tazzina da caffè ha un solo piano di simmetria, la Grande Piramide ne ha quattro. Un cubo ne ha nove: tre sono paralleli a una coppia di facce opposte e sei passano per le diagonali corrispondenti di due facce opposte; un cilindro e una sfera hanno un numero infinito di piani di simmetria.

Pensate a un oggetto solido diviso in due da un piano di simmetria; se mettete una delle due metà di fronte a uno specchio ponendo contro il vetro la sezione trasversale, l'immagine nello specchio insieme alla metà postavi davanti ricomporranno la forma originaria. Qualsiasi solido con almeno un piano di simmetria si può sovrapporre alla sua immagine speculare facendogli compiere, se necessario, una rotazione nello spazio.

Parlando di questo argomento nel mio libro *The Ambidextrous Universe* (Charles Scribner's Sons, 1979) a pagina 19 affermai che, se un oggetto tridimensionale non ha nessun piano di simmetria (come una spirale, una striscia di Möbius, un nodo semplice in un anello di corda), non si può sovrapporre alla sua immagine speculare senza immaginare di fargli compiere una rotazione impossibile che lo rovesci in una quarta dimensione.

È falso! Come molti lettori del libro hanno precisato, ci sono figure solide totalmente prive di un piano di simmetria che possono ugualmente essere sovrapposte alle loro immagini speculari mediante un'adeguata rotazione nello spazio ordinario. Una di queste sovrapposizioni è così semplice che la potreste fare in un batter d'occhio piegando un foglio di carta quadrato. Come si fa?

3. La coperta patchwork rovinata. La coperta della figura in alto della pagina a fronte misura 9×12 ed era originariamente composta di 108 quadrati unità; il consumo rese necessario sostituire otto quadrati al centro come è indicato nella figura.

Il problema è il seguente: tagliate la coperta lungo le linee in due soli pezzi che si possano cucire insieme formando una coperta quadrata 10 × 10. La nuova coperta non può, ovviamente, avere buchi. Ciascuna parte può venir ruotata, ma non rovesciata perché il rovescio della coperta non coincide con il diritto.

Il rompicapo è vecchio, ma la sua soluzione è così bella e il problema è così poco noto che io ricevo continuamente lettere dai lettori che non sono a conoscenza della sua origine. La soluzione è unica anche se non si richiede che il taglio venga fatto lungo le linee del reticolo.

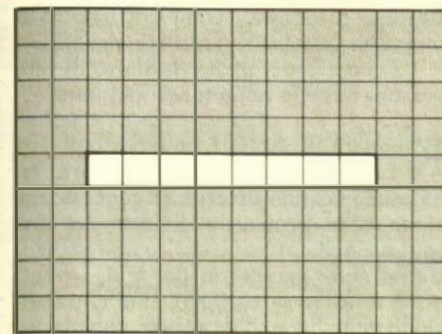
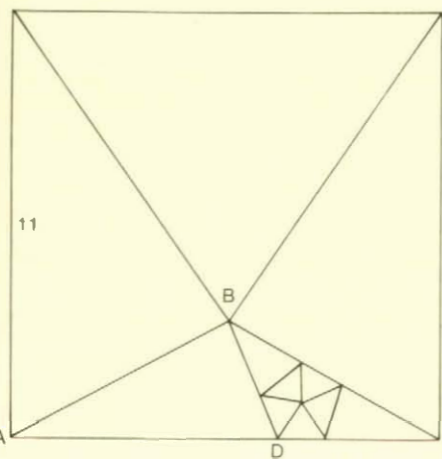
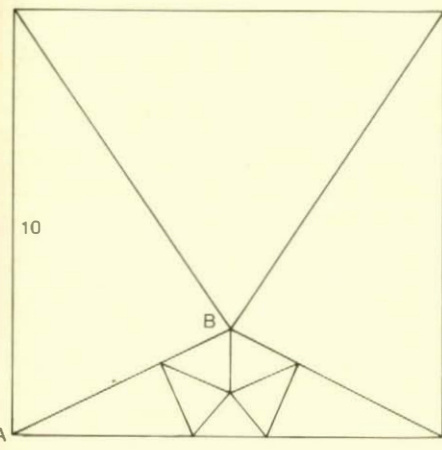
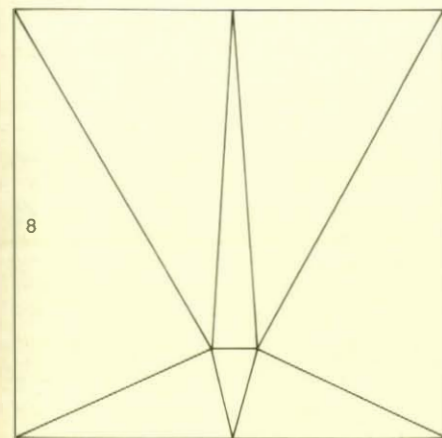
4. Triangoli acuti e isosceli. Un triangolo è acuto quando tutti i suoi angoli interni sono inferiori a 90 gradi. Qual è il minor numero di triangoli acuti che non si sovrappongano in cui si può dividere un quadrato?

Mi sono posto il problema circa 20 anni fa e l'ho risolto dimostrando come fosse possibile dividere un quadrato in 8 triangoli acuti alla maniera indicata nella prima figura dell'illustrazione della pagina a fronte. Riferendone nella rubrica, ristampata poi come capitolo 3 del mio *New Mathemathical Diversions from Scientific American* (Simon and Schuster, 1966), dicevo: «Da giorni ero convinto che la risposta fosse nove, quando improvvisamente vidi come ridurre i triangoli a otto.»

Da allora ho ricevuto molte lettere dai lettori che non riuscivano a trovare una soluzione con 9 triangoli acuti, ma puntualizzavano che erano possibili soluzioni per 10 o qualsiasi numero superiore di triangoli. La seconda figura dell'illustrazione della pagina a fronte mostra una soluzione con 10 triangoli. Si noti che il triangolo ottuso *ABC* è diviso in sette triangoli acuti da un pentagono di cinque triangoli acuti. Se ora *ABC* viene diviso da *BD*, come è indicato nella figura in basso, in due triangoli uno acuto e uno ottuso, possiamo servirci sempre del metodo del pentagono per dividere il triangolo acuto *BCD* in sette triangoli acuti ottenendo quindi per l'intero quadrato 11 triangoli acuti. Ripetendo il procedimento si otterranno 12, 13, 14... triangoli acuti.

Di fatto il problema più difficile è quello di dividere in nove triangoli acuti, ma, nonostante tutto, si può fare come rivelerò nella mia prossima rubrica.

Ci sono molti problemi analoghi che chiedono di dividere figure in triangoli che non si sovrappongono; io ne citerò solo due. È facile dividere un quadrato in un numero pari di triangoli di area uguale, ma la cosa si può fare con un numero dispari di triangoli? La sorprendente risposta è no. Per quanto mi risulta, la cosa è stata dimostrata da Paul Monsky per primo sull'«American Mathematical Monthly» (Vol. 77, n. 2, pagg. 161-164; febbraio 1970).

Il Viale dell'Università dove Lavinia deve trovare una stanza

La coperta patchwork rovinata

Divisione di un quadrato in triangoli acuti

Un altro curioso teorema è che qualsiasi triangolo si può dividere in n triangoli isosceli purché n sia maggiore di 3. Lo ha dimostrato Gali Salvatore su «Crux Mathematicorum» (Vol. 3, no. 5, pagg. 134-135; maggio 1977). È di particolare interesse il caso del triangolo equilatero: è facile dividerlo in quattro triangoli isosceli (tutti equilateri) o in tre triangoli isosceli. (Alcuni triangoli non possono essere divisi in due o tre triangoli isosceli; ecco perché il teorema richiede che n sia 4 o un numero maggiore di 4.) Siete capaci di dividere un triangolo equilatero in cinque triangoli isosceli? Nella mia prossima rubrica farò vedere come si può fare senza che nessuno dei cinque triangoli sia equilatero, con un solo triangolo equilatero e con due soli triangoli equilateri su cinque. Non è possibile che più di due dei triangoli isosceli siano equilateri.

5. Misurare con lo yen. Mi ha dato lo spunto per questo problema Mitsunobu Matsuyama, un lettore di Tokio che mi mandò una scorta di monete giapponesi da uno yen e me ne descrisse le seguenti notevoli proprietà poco note anche in Giappone. La moneta da uno yen è fatta di puro alluminio, ha un raggio di esattamente un centimetro e pesa esattamente un grammo. Gli yen possono quindi essere usati su una bilancia per determinare il peso in grammi di piccoli oggetti. Si può anche usare su una superficie piana per misurare in centimetri le distanze.

È facile rendersi conto di come monete da uno yen si possono disporre su di una linea per misurare distanze pari (due centimetri, quattro centimetri, sei e così via), ma possono servire anche per misurare distanze dispari (uno, tre, cinque centimetri e così via)? Dimostrare come le monete da uno yen possano servire per misurare in centimetri tutte le distanze intere lungo una retta.

6. Un nuovo gioco sulla colorazione delle mappe. L'ideatore di questo problema è Steven J. Brams, un politologo della New York University che è l'autore di *Game Theory and Politics* (1975) e *Paradoxes in Politcs (1976)*, entrambi pubblicati dalla Free Press, di *The Presidential Election Game* (Yale University Press, 1978). Il suo ultimo libro *Biblical Games* (The MIT Press, 1980) è una sorprendente applicazione della teoria dei giochi a quegli episodi dell'Antico Testamento traducibili in un gioco in cui si suppone che uno dei giocatori sia una divinità onnisciente.

Supponete di avere una mappa piana finita e connessa e di disporre di un numero n di pastelli di diverso colore. Il primo giocatore, il minimizzatore, sceglie un qualsiasi pastello e colora una qualsiasi regione della mappa. Il secondo giocatore, il massimizzatore, a sua volta, colora una qualsiasi altra regione usando uno degli n colori. I giocatori continuano così colorando a turno una regione con uno degli n colori obbedendo sempre alla regola che due regioni dello stesso colore non possono avere un confine comune. Colori uguali possono, ovviamente, toccarsi per punti.

Il minimizzatore cerca di evitare che vi sia bisogno di un $n + 1$-esimo colore per completare la mappa; il massimizzatore cerca di obbligarne l'uso. Il massimizzatore vince se uno dei giocatori non può più giocare servendosi di uno degli n pastelli prima che la mappa sia stata interamente colorata; se invece l'intera mappa viene colorata con n colori è il minimizzatore a vincere.

Il sottile e difficile problema è: qual è il valore minimo di n tale per cui, quando il gioco viene giocato su una qualsiasi mappa, il minimizzatore può sempre vincere, posto che entrambi i giocatori giochino in maniera ottimale?

Per rendere più chiaro il problema si consideri la semplice mappa dell'illustrazione in alto della pagina seguente. Essa dimostra che n deve essere almeno 5. Ovviamente se non si fa il gioco, la mappa può essere facilmente colorata con 4 colori come qualsiasi mappa del piano (è il famoso teorema dei quattro colori che ora si sa essere vero), ma se si fa il gioco di Brams e sono disponibili solo 4 colori, il massimizzatore può sempre costringere il minimizzatore a usare un quinto colore. Se sono disponibili 5 colori, il minimizzatore può sempre vincere.

Brams ipotizza che il valore minimo di n sia 6. È stata però trovata una mappa sulla quale con cinque colori il massimizzatore può sempre vincere. Siete in grado

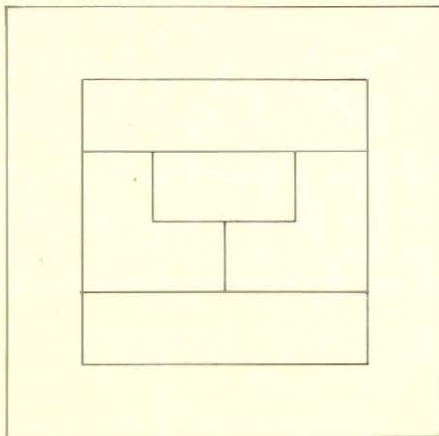
ERRATA CORRIGE

Ci scusiamo con i nostri lettori per una trasposizione di testi avvenuta involontariamente alle pagine 138, 139 e 140 del fascicolo del mese scorso nella rubrica «Temi metamagici».

Per una corretta lettura dell'articolo si tenga presente quanto segue:

dalla 13ª riga dal fondo della 3ª colonna di pagina 138 si deve passare alla pagina 139;

le ultime 11 righe di pagina 138 vanno spostate in calce a pagina 139;

la 12ª riga dal fondo di pagina 138 va inserita tra la 10ª e la 9ª riga dal fondo della prima colonna di pagina 140.

Il gioco di Brams della colorazione della mappa

di costruire tale mappa e di indicare la strategia vincente del massimizzatore? Ricordatevi che il minimizzatore gioca per primo e che nessun giocatore è obbligato a introdurre un nuovo colore a ogni turno se può fare un gioco consentito usando un colore di cui ci si è già serviti.

7. Whim. Nel suo libro *Gödel, Escher, Bach* (ora in edizione economica) Douglas R. Hofstadter introduce il concetto di giochi automodificantisi che sono giochi in cui è permesso a un giocatore, quando è il suo turno, di annunciare una nuova regola che modifica il gioco. La nuova regola viene chiamata metaregola e una regola che modifica una metaregola si dice metametaregola e così via. Hofstadter dà alcuni esempi chiave. Invece di muovere un giocatore potrebbe annunciare che da quel momento in avanti non si può più occupare una certa casella o che tutti i cavalli devono muoversi in un modo un po' differente o qualsiasi altra metaregola che faccia parte di un elenco di alterazioni consentite del gioco.

L'idea base non è interamente nuova; prima del 1970 John Horton Conway propose una bizzarra variazione automodificantesi del nim che chiamò whim. Il nim è un gioco per due persone che si gioca con gettoni che sono disposti in un numero arbitrario di mucchi con un numero arbitrario di gettoni ciascuno. I giocatori, a turno, tolgono uno o più gettoni da un qualsiasi mucchio. Nel nim normale vince la persona che prende l'ultimo gettone, nella versione misère la persona che prende l'ultimo gettone perde. La strategia per il gioco perfetto si conosce da molto tempo e voi potete trovarla nel capitolo sul nim del mio *Scientific American Book of Mathematical Puzzles and Diversions* (Simon and Schuster, 1959).

Il Whim inizia senza che sia stato deciso se il gioco sarà normale o misère. In qualsiasi momento del gioco tuttavia un giocatore può, invece di muovere, annunciare se il gioco è normale o misère. Questa «mossa del whim» viene fatta una volta sola e da allora in avanti il gioco è deciso. È ben noto che nel nim la strategia è la stessa per entrambe le forme fin quasi alla fine; potreste quindi essere tentati di supporre che la strategia del whim si analizzi facilmente. Provate a giocare qualche partita e vedrete che non è così semplice come sembra!

Supponete di essere il primo a giocare in una partita con molti mucchi e molti gettoni in ciascun mucchio e che la vostra posizione vi porterebbe alla sconfitta nel nim. Dovreste subito fare la mossa del whim perché lascia le posizioni invariate ma vi fa diventare il giocatore vincente. Supponete tuttavia di essere il primo a giocare e che la vostra posizione vi porti alla vittoria. Non osate fare la mossa vincente perché questo permette al vostro avversario di dichiarare il whim e di mettervi in posizione perdente. Dovete quindi fare una mossa che sarebbe perdente nel nim normale e, per la stessa ragione il vostro avversario dovrebbe riponderc con una mossa perdente. Ovviamente, se un giocatore non fa una mossa perdente, l'altro vince dichiarando il whim.

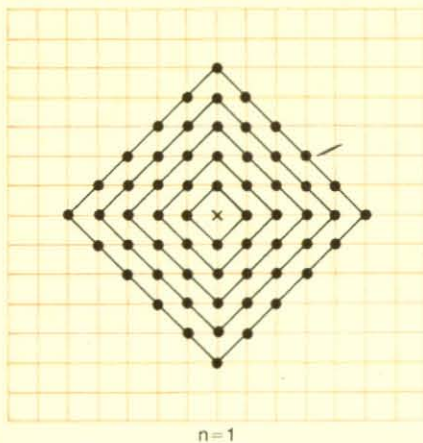
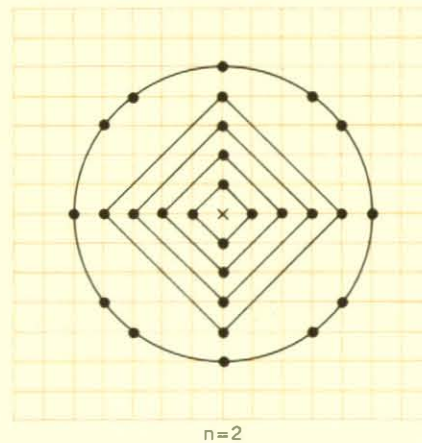
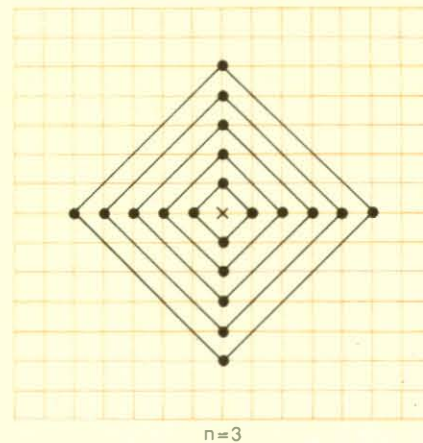
Quando il gioco si avvicina alla fine raggiungendo il punto in cui la strategia vincente diverge per le due forme di nim, può essere necessario dichiarare il whim per vincere. Come si può determinare se farlo o non farlo? E come si può decidere all'inizio della partita chi vincerà se entrambi i giocatori giocano al meglio? La strategia di Conway, come egli ha notato una volta, è facile da ricordare, ma difficile da indovinare anche da parte di una persona versata nella teoria del nim.

Kenneth W. Abbott, un esperto di teoria dei calcolatori di New York, mi ha mandato una divertente generalizzazione della geometria discreta del taxi trattata nella rubrica del marzo scorso. Come nella geometria del taxi, i punti dello spazio non euclideo sono l'intersezione delle linee di un reticolo quadrato. Nella generalizzazione di Abbott la «distanza» tra due punti qualsiasi è un intero definito come uguale a $\sqrt[n]{x^n+y^n}$, dove x è misurato orizzontalmente, y è misurato verticalmente e n è un intero positivo.

Quando n è uguale a 1, abbiamo la geometria del taxi semplice descritta in marzo. Tutti i «cerchi» di questa geometria sono insiemi di punti equidistanti dal centro del cerchio e hanno la forma che si vede nella figura in basso, dove i raggi sono 1, 2, 3, 4 e 5.

Quando n è uguale a 2, i cerchi con lo stesso raggio prendono la forma che si vede nell'illustrazione in mezzo. Si noti che i primi quattro cerchi consistono di esattamente 4 punti che giacciono sui due assi che passano per il centro comune ai cinque cerchi. Definiremo tali cerchi «banali». Quando n è uguale a 1, solo il cerchio di raggio 1 è banale, mentre tutti gli altri cerchi non lo sono. Quando n è uguale a 2, il quinto cerchio è non banale. In questa geometria ci sono un numero infinito di entrambi i tipi di cerchi. π è $2\sqrt{2}$ per tutti i cerchi banali, ma ha valori differenti per quelli non banali. Per il quinto cerchio che ha raggio 5, π è $(4\sqrt{10}+2\sqrt{2})/5$.

Quando n è uguale a 3, i primi cinque cerchi *(a destra nella figura in basso)* sono tutti banali. In questa geometria π è uguale a $2^{(n+1)/n}$ per tutti i cerchi banali.

Facciamo ora un'importante ipotesi. Qualsiasi geometria dei taxi generalizzata con n maggiore di 2 può contenere solo cerchi banali. Questa ipotesi non dimostrata, come ha precisato Abbott, risulta facilmente equivalente all'ultimo teorema di Fermat.

«Cerchi» di raggio 1, 2, 3, 4 e 5 in una geometria generalizzata del «taxi»

Sommario